Shahid Rajaee
Teacher Training University

Journal of
Electrical and Computer
Engineering Innovations
(JECEI)

Vol. 12 No. 1, Winter-Spring 2024

Journal of Electrical and Computer Engineering Innovations Vol. 12 No. 1 Winter-Spring 2024

JECEI

Journal of

Shahid Rajaee
Teacher Training University

*p*ISSN 2322-3952
*e*ISSN 2345-3044

Electrical and Computer
Engineering Innovations (JECEI)

Semiannual Publication

Volume 12, Issue 1, Winter-Spring 2024

Journal of Electrical and Computer Engineering Innovations
http://jecei.sru.ac.ir
EISSN: 2345-3044 ISSN: 2322-3952
JECEI
SRTTU
JECEI

License Holder: Shahid Rajaee Teacher Training University (SRTTU)

Address: Lavizan, 16788-15811, Tehran, Iran.

Journal of Electrical and Computer Engineering Innovations

Vol. 12; Issue 1: 2024

Contents

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Clustering of Triangular Fuzzy Data Based on Heuristic Methods

N. Ghanbari [1], S. H. Zahiri [1,*], H. Shahraki [2]

[1]Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

[2]Department of Computer Engineering, Faculty of Industry and Mining, University of Sistan and Baluchestan, Khash, Iran.

Article Info

Article History:
Received 18 March 2023
Reviewed 23 April 2023
Revised 16 May 2023
Accepted 14 June 2023

Keywords:
Heuristic clustering
Particle swarm optimization
Uncertain data
Fuzzy dataset
Ranking function

*Corresponding Author's Email Address: *hzahiri@birjand.ac.ir*

Abstract

Background and Objectives: In this paper, a new version of the particle swarm optimization (PSO) algorithm using a linear ranking function is proposed for clustering uncertain data. In the proposed Uncertain Particle Swarm Clustering method, called UPSC method, triangular fuzzy numbers (TFNs) are used to represent uncertain data. Triangular fuzzy numbers are a good type of fuzzy numbers and have many applications in the real world.

Methods: In the UPSC method input data are fuzzy numbers. Therefore, to upgrade the standard version of PSO, calculating the distance between the fuzzy numbers is necessary. For this purpose, a linear ranking function is applied in the fitness function of the PSO algorithm to describe the distance between fuzzy vectors.

Results: The performance of the UPSC is tested on six artificial and nine benchmark datasets. The features of these datasets are represented by TFNs.

Conclusion: The experimental results on fuzzy artificial datasets show that the proposed clustering method (UPSC) can cluster fuzzy datasets like or superior to other standard uncertain data clustering methods such as Uncertain K-Means Clustering (UK-means) and Uncertain K-Medoids Clustering (UK-medoids) algorithms. Also, the experimental results on fuzzy benchmark datasets demonstrate that in all datasets except Libras, the UPSC method provides better results in accuracy when compared to other methods. For example, in iris data, the clustering accuracy has increased by 2.67% compared to the UK-means method. In the case of wine data, the accuracy increased with the UPSC method is 1.69%. As another example, it can be said that the increase in accuracy for abalone data was 4%. Comparing the results with the rand index (RI) also shows the superiority of the proposed clustering method.

Introduction

Data clustering plays an essential role in machine learning, data mining, statistical analysis, image segmentation, and pattern recognition [1]-[3]. Clustering groups data objects based on the information found in that qualifies the objects and their communications. The aim of clustering is that the objects within a group be similar or related to one another and different from the objects in the other groups.

The greater likeness within a group and the greater disagreement between the groups, the better the clustering [4]. Almost all standard heuristic clustering methods are developed for crisp datasets. At the same time, in real-world applications such as sensor measurements, biomedical, and microarray, information cannot often be expressed by crisp numbers. For simple example in the sentence "Ryan's intelligence is good." The goodness of Ryan's intelligence is a qualitative sense and

cannot be measured by any quantity. These data are known as uncertain data.

Uncertain data can be used in many cases. There are some quantities that must be mentioned by interval data; For instance, the range of temperature during a day. Also, uncertainty may have resulted from the lack of knowledge, implicit randomness in data generation, data staling, and inability to perform adequate physical measurements or vagueness [5]. It is usually related to missing or incomplete information or the probability of occurrence of given information [6]. Therefore, it is a critical issue to consider uncertain data in clustering algorithms. The clustering of uncertain data has been considered in various papers [7]-[13]. In some studies, interval data is used to represent uncertain data [7]. A famous example of interval data is the temperature data set [7]. The temperature interval data set gives the minimum and the maximum monthly temperatures of 37 cities in degrees centigrade. This data set has 12 features and 4 cluster. In [8], an uncertain clustering algorithm using cloud model is proposed. In [9], one method of multiple kernel fuzzy clustering for uncertain data classification has been presented. A suitable cluster validity index for uncertain data clustering is presented in [10]. In [11], a new clustering algorithm is presented using the k-medoids for uncertain objects. In [12], uncertain data clustering in distributed peer-to-peer networks has been done. In [13], an upgrade of particle swarm optimization (PSO) algorithm for the fuzzy environment has been reported where particles have been defined as fuzzy numbers, and their motion has been reformulated by fuzzy equations.

In [14], interval data clustering based on the adaptive dynamic cluster method has been described. Carvalho et al. used adaptive Hausdorff distances and dynamic clustering to cluster the interval data [15]. In [16], a novel clustering method based on novel density and hierarchical density has been proposed for interval data. Likewise, using interval data in various methods and applications is described in [7]. Using fuzzy numbers is another common way to express uncertain data. Tayyebi proposed a fuzzy clustering method (FCM), which clusters trapezoidal fuzzy numbers [17]. In this method, a linear ranking function is applied to define a distance between trapezoidal fuzzy data.

On the other hand, the powerfulness of optimization algorithms has caused those the heuristic clustering methods such as the genetic algorithm-based clustering method [18], [19] particle swarm optimization algorithm-based clustering method [20], [21], learning automata-based clustering method [22], and clustering method based on gravitational search algorithm [23], [24] be created. In these clustering methods, the optimization algorithms are used to find the optimal centroids for clusters.

This type of clustering method (heuristic clustering method) is used in many types of research [18]-[24], and their results show that these clustering methods have good potential to apply in different applications. Despite all the advances of these clustering methods [18]-[24], clustering of uncertain data with these methods is one of the topics that have not yet been investigated. So, in this paper, we want to promote one of these clustering methods to cluster uncertain data. Thus, the particle swarm optimization algorithm-based clustering method, which is one of the best clustering methods of this family, is chosed. The particle swarm optimization (PSO) algorithm is a relatively novel heuristic algorithm introduced by Kennedy and Eberhart [25]. PSO algorithm efforts to find the optimal solution throughout the simulation of some concepts that obtained from bird flocking, fish schooling, and other social folks. Each particle can efficiently attain his objective using the information that is owned by itself and the information that is assigned between the folk. This means that the PSO algorithm is an optimization procedure that utilizes the laws of social behavior [26]. Data clustering using the PSO algorithm was first suggested by Engelbrecht and Merwe in [27]. The idea of this clustering algorithm was allocating all cluster centroids to each particle and update the particle following to the fitness function value computed for the particle. As mentioned, particle swarm optimization has used only for certain data until now. In this paper, an enhancement of the PSO algorithm is proposed to appropriate for clustering uncertain data (triangular fuzzy numbers).

In this paper, for the first time, the new version of the particle swarm optimization algorithm is proposed for clustering uncertain data. For presenting uncertain data, triangular fuzzy numbers are used. On the other hand, in the present paper, particle swarm optimization algorithm is promoted to proceed with clustering triangular fuzzy numbers. The proposed method is called the Uncertain Particle Swarm Clustering (UPSC) method. In the UPSC method, particles are triangular fuzzy numbers. Therefore, to upgrade a version of standard PSO, obtaining the distance between the fuzzy numbers is necessary. For this purpose, a linear ranking function is applied to describe the distance between fuzzy vectors.

Ranking function used in the proposed method (UPSC) is introduced for triangular fuzzy numbers; because triangular fuzzy numbers are a good type of fuzzy numbers and have many applications in the real world. Much of the data in the real world is expressed with specific precision, which can be well represented by a triangular fuzzy number. The proposed method can also be used for clustering datasets of the type of interval data, real data or trapezoidal fuzzy number.

The contributions of this paper are as follows:

1- The method proposed in this paper can cluster uncertain or fuzzy data. In other words, uncertain data can be given as input to the proposed algorithm and be clustered.
2- in order to capable cluster uncertain data by the proposed algorithm, it is necessary to make changes in the conventional PSO algorithm. An important change is the use of a linear ranking function in the fitness function of PSO algorithm. With this, the distance between fuzzy/uncertain numbers can be calculated and evaluated. The changes applied to the conventional PSO algorithm provide a new version of the conventional PSO algorithm that can cluster fuzzy/uncertain data.
3- The new proposed cost function (using linear ranking function) of the proposed method can find the optimal centers of the clusters for data whose features are expressed in triangular fuzzy numbers (TFNs).
4- The most important advantage of the proposed method is compatibility with real-life applications. For example, the proposed method is its applicability in missing value data because one way to express missing value data is to estimate it with triangular fuzzy numbers by using the Fuzzy Nearest Neighborhood Mean (FNNM) method [17].
5- The proposed method can also be implemented for crisp data because crisp data is a particular type of fuzzy data and can be represented as triangular fuzzy data easily. It is necessary to mention uncertain data is uncertain due to noise. The fact that we defined a fuzzy membership function for each input data indicates that it is noisy.
6- One of the strengths of the proposed method is its general and modular form. Most of the computational work of the proposed method, such as fitness function and fuzzification of features, are general, and other heuristic methods with minimum changes can be utilized instead of PSO.

The rest of this paper are constituted as follows: Initially some preliminary notations and concepts of fuzzy numbers and fuzzy theory is presented. Then the proposed particle swarm optimization algorithm for clustering fuzzy data and the experimental clustering results are reported. Finally, conclusion of the paper is reported.

Preliminary Concepts

This section is divided into two subsections. Subsection "A" describes several concepts of the fuzzy set theory. Due to the importance of ranking function in the method presented in the present paper, Subsection "B" will investigate this issue, and in it, a linear ranking function for comparing fuzzy numbers are introduced. While introducing ranking function, the necessary definitions and lemmas is provided.

A. A Brief Introduction to Fuzzy Set Theory

In this section, we review the main concepts of the fuzzy set theory, initialize by Bellman and Zadeh in [28], and is applied in this paper.

Let UNI be the universal set. A mapping $\widetilde{\mathrm{P}}: \mathrm{UNI} \rightarrow [0,1]$ is a fuzzy set. The value $\tilde{P}(x)$ of $\widetilde{\mathrm{P}}$ at $x \in UNI$ stands for the grade of membership of x in $\widetilde{\mathrm{P}}$. A fuzzy set $\widetilde{\mathrm{P}}$ is normal if there exists $x_0 \in UNI$ such that $P(x_0) = 1$. An α-cut of fuzzy number $\widetilde{\mathrm{P}}$, $\alpha \in [0,1]$, is a crisp set as:

$$\alpha = \{x \in \mathrm{UNI}: \widetilde{\mathrm{P}}(x) \geq \alpha\} \quad (1)$$

If a fuzzy set $\widetilde{\mathrm{P}}$ satisfies that $\widetilde{\mathrm{P}}_\alpha$ is a closed interval for every $\alpha \in [0,1]$, then $\widetilde{\mathrm{P}}$ is called a fuzzy number. A particular type of fuzzy numbers is the triangular fuzzy number (TFN) to be determined as:

$$\tilde{P}_\alpha = \begin{cases} 0 & \text{for} \quad x \leq P_1 \\ \frac{x - P_1}{P_2 - P_1} & \text{for} \quad P_1 < x \leq P_2 \\ \frac{P_3 - x}{P_3 - P_2} & \text{for} \quad P_2 < x \leq P_3 \\ 0 & \text{for} \quad P_3 < x \end{cases} \quad (2)$$

For reduction, the TFN $\widetilde{\mathrm{P}}$ has been marked by $\widetilde{\mathrm{P}} = (\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3)$ (Fig. 1). Next, arithmetic operation on triangular fuzzy numbers is described. Suppose that $\tilde{\mathrm{p}} = (\mathrm{p}_1, \mathrm{p}_2, \mathrm{p}_3)$ and $\tilde{\mathrm{q}} = (\mathrm{q}_1, \mathrm{q}_2, \mathrm{q}_3)$ be two triangular fuzzy numbers and c a real number. Scalar product and scalar addition operators are described as follows:

$$c.\tilde{\mathrm{p}} = (c\mathrm{p}_1, c\mathrm{p}_2, c\mathrm{p}_3) \quad \text{if } c \geq 0, c \in \mathrm{R} \quad (3)$$

$$c.\tilde{\mathrm{p}} = (c\mathrm{p}_3, c\mathrm{p}_2, c\mathrm{p}_1) \quad \text{if } c \leq 0, c \in \mathrm{R} \quad (4)$$

$$\tilde{\mathrm{p}} + \tilde{\mathrm{q}} = (\mathrm{p}_1 + \mathrm{q}_1, \mathrm{p}_2 + \mathrm{q}_2, \mathrm{p}_3 + \mathrm{q}_3) \quad (5)$$

$$\tilde{\mathrm{p}} - \tilde{\mathrm{q}} = (\mathrm{p}_1 - \mathrm{q}_3, \mathrm{p}_2 - \mathrm{q}_2, \mathrm{p}_3 - \mathrm{q}_1) \quad (6)$$

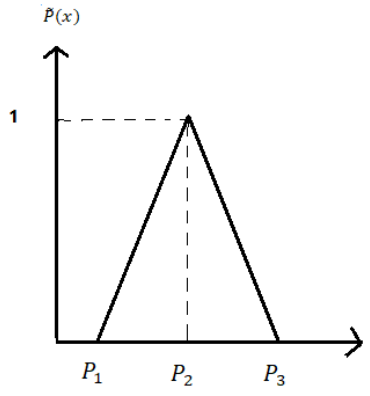


Fig. 1: Membership function of TFN $\widetilde{\mathrm{P}} = (\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3)$.

B. Ranking Function

There are several procedures for comparing fuzzy numbers, which can be viewed in Tanaka and Ichihashi [29], Lai and Hwang [30], Fang and Hu [31], and Shoacheng [32].

One of the most suitable of these procedures has

relied on the notion of comparison of fuzzy numbers using ranking functions [33]-[41]. In this section, a ranking function have been investigated to determine a distance between fuzzy vectors.

An effective approach for arranging the point of F(R) is to determine a ranking function R: F(R)→R that maps each triangular fuzzy number into the actual line, where a natural order exists.

The ranking function is applied to specify a distance between triangular fuzzy vectors in the section 3. In this section, only linear ranking function are considered, i.e., a ranking function R such that

$$R(\tilde{p} + c\tilde{q}) = R(\tilde{p}) + cR(\tilde{q}) \quad (7)$$

for any $\tilde{p}$, $\tilde{q}$∈F(R) and any c ∈R, it is clear that R ($\tilde{0}$) = 0, where $\tilde{0}$ = (0,0,0). The ranking function is defined using Lemma.

Lemma 2.2.1 For constant nonnegative numbers α, β∈R, the function R: F(R)→R is determined as:

$$R(\tilde{p}) = \alpha p_1 + 2\beta p_2 + \alpha p_3 \quad (8)$$

where $\tilde{p} = (p_1, p_2, p_3)$∈F(R) is a linear ranking function. The proof was given in [17]. For example, if α=β=1/4, then $R(\tilde{p}) = p_1 + 2p_2 + p_3/4$, that has been suggested by Yager [42].

In the proposed method (UPSC), we set α=β=1/4 for the ranking function.

Then, for triangular fuzzy numbers $\tilde{p}$ and $\tilde{q}$, we have

$$\tilde{p} \geq \tilde{q} \quad \text{if and only if } p_2 + \frac{1}{2}(p_3 + p_1) \geq q_2 + \frac{1}{2}(q_3 + q_1) \quad (9)$$

Ranking function used in the proposed method (UPSC) are introduced for triangular fuzzy numbers; because triangular fuzzy numbers are a standard approximation for fuzzy numbers and have many applications in the real world. Much of the data in the real world is expressed with specific precision, which can be well represented by a triangular fuzzy number. For example, persons 'weight can be defined with a sure accuracy, which can be well represented by triangular fuzzy numbers. The proposed method (UPSC) can also be applied for clustering datasets of the type of interval data, real data or trapezoidal fuzzy number.

A ranking function to specify a distance between trapezoidal fuzzy vectors/numbers is explained in [17]. Obviously, any real number and any interval number can be rewritten as a trapezoidal fuzzy number.

Uncertain Particle Swarm Clustering (UPSC) Method

In this section, at first, an overview of PSO and then PSO clustering method for certain data and finally proposed method (UPSC) for uncertain data will be explained respectively.

C. An Overview of PSO

The particle swarm optimization (PSO) algorithm is a relatively novel heuristic algorithm introduced by Kennedy and Eberhart [25], [26]. In PSO method, the particles fly through the problem/search space by following the optimal particles. Each particle recollects the best position that it has searched (P_{best}) and also best position among all the particles in the Population/group (G_{best}). The position of each particle updates according to the P_{best} and G_{best} in the search space. A simple and standard implementation of PSO algorithm is as follows [43]:

1. Create a random population of particles in the search space
2. Calculate the fitness of each particle according to the fitness function defined in the problem
3. Update the velocity of each particle based on the velocity (10)

$$vel_i^d(t+1) = Wvel_i^d(t) + rand \times c_1 \times \left[p_{best_i}^d - x_i^d(t)\right] + rand \times c_2 \times \left[g_{best_i}^d - x_i^d(t)\right] \quad (10)$$

where, c_1 and c_2 are two positive constants, rand is random number uniformly distributed within the span [0,1], w is the inertia weight. In the proposed method (UPSC), we set c_1=c_2=2 and w=0.7. position of the ith particle shows with $X_i = (x_i^1, x_i^2, \dots, x_i^n)$ and velocity of the ith particle shows with $VEL_i = (vel_i^1, vel_i^2, \dots, vel_i^n)$. Also, the best previous position of the ith particle represent with $P_{besti} = (p_{best_i}^1, p_{best_i}^2, \dots, p_{best_i}^n)$ and the best previous position among all the particles in the group shows with $G_{besti} = (g_{best_i}^1, g_{best_i}^2, \dots, g_{best_i}^n)$.

4. Update the position of each particle using the position (11)

$$x_i^d(t+1) = vel_i^d(t+1) + x_i^d(t) \quad (11)$$

5. Repeat loops 2 to 4 until the stop condition is met. The condition for stopping is usually to achieve the desired accuracy or maximum number of repetitions.

D. PSO Clustering Method for Certain Data

Among all the literature attempts to improve the particle swarm optimization algorithm to real/certain data clustering, [44] represents to be the one that is nearest to the main ideas of the PSO because each particle perceives an entire candidate solution to the problem. In other words, each particle represents the center of the cluster. Based on this method, a particle X_i is created as follows:

$$X_i = (m_{i1}, m_{i2}, \dots, m_{ic}) \quad \text{for} \quad L=1,2,\dots,c \quad (12)$$

For clusters with n dimension m_{iL} is as follows.

$$m_{iL} = (m_{iL}^1, m_{iL}^2, \dots, m_{iL}^n) \quad \text{for} \quad p=1, 2,\dots,n \quad (13)$$

where c is the number of clusters to be organized, and

m_{iL} corresponds to the Lth centroid of the ith particle, the centroid of the cluster C_{iL}. So, a lonely particle demonstrates a candidate solution to a given clustering problem.

The input data set D contains N samples with dimensions n.

$$D = (d_1, d_2, \dots, d_N) \quad \text{for } k = 1,2,3,\dots,N \tag{14}$$

where d_k defines as follows:

$$d_k = (d_k^1, d_k^2, \dots, d_k^n) \quad \text{for} \quad p = 1,2,3,\dots,n \tag{15}$$

The fitness of each particle is evaluated by the following fitness function.

$$\text{Fitness function} = \frac{\sum_{L=1}^{c}[\sum_{\forall d_k \in C_{iL}} ED(d_k, m_{iL})/|C_{iL}|]}{c} \tag{16}$$

where $d_k = (d_k^1, d_k^2, \dots, d_k^n)$ defines the kth data vector, $|C_{iL}|$ is the number of data vectors belonging to the cluster C_{iL} and ED is the Euclidian distance between d_k and m_{iL}.

E. The Proposed Uncertain Particle Swarm Clustering (UPSC) Method

In this section, a novel PSO algorithm is offered for clustering triangular fuzzy data. This algorithm expands the standard PSO algorithm introduced in [44] because the standard PSO algorithm just clusters certain/real data. Suppose we have a dataset such as:

$$\tilde{D} = [\tilde{d}_1, \tilde{d}_2, \dots., \tilde{d}_N] \quad \text{for } k = 1,2,3,\dots,N \tag{17}$$

where N is the number of data vector to be clustered.

In (17) we have:

$$\tilde{d}_k = (\tilde{d}_k^1, \tilde{d}_k^2 \dots, \tilde{d}_k^n)^T \quad \text{for} \quad p = 1,2,3,\dots,n \tag{18}$$

And

$$\tilde{d}_k^p = ((d_k^p)_1, (d_k^p)_2, (d_k^p)_3) \quad \text{for } k = 1,2,3,\dots,N \quad \text{and} \quad p = 1,2,3,\dots,n \tag{19}$$

Thus, $\tilde{D}$ is a part of $F^{n\times N}(R)$. The goal is to divide $\tilde{d}_k$'s into *c* clusters.

Allow $\tilde{X}_i$ be a triangular fuzzy matrix of the prototype. $\tilde{X}_i$ is defined as:

$$\tilde{X}_i = (\tilde{m}_{i1}, \tilde{m}_{i2}, \dots, \tilde{m}_{ic}) \quad \text{for} \quad L=1,2,\dots,c \tag{20}$$

$\tilde{X}_i$ is a part of $F^{n\times c}(R)$. In (20) we have:

$$\tilde{m}_{iL} = (\tilde{m}_{iL}^1, \tilde{m}_{iL}^2, \dots, \tilde{m}_{iL}^n) \quad \text{for} \quad p=1, 2,\dots,n \tag{21}$$

$$\tilde{m}_{iL}^p = ((m_{iL}^p)_1, (m_{iL}^p)_2, (m_{iL}^p)_3) \quad \text{for } L = 1,2,3,\dots,c \quad \text{and} \quad p = 1,2,3,\dots,n \tag{22}$$

Because the data are fuzzy, it is assumed that the prototype is also triangular fuzzy numbers. But in the implementation of the proposed method (UPSC), it was decided to obtain cluster centers as real numbers, so in (22) $(m_{iL}^p)_1 = (m_{iL}^p)_2 = (m_{iL}^p)_3$ is considered. In the following, the linear ranking function is applied to describe a distance between the fuzzy vector $\tilde{d}_k$'s and vector $\tilde{m}_{iL}$'s to be necessary to extend the standard PSO algorithm.

Definition 3.3.1 allow R be a linear ranking function. The mapping $d_R: F^n(R) \times F^n(R) \to R$ with

$$d_R(\tilde{f}, \tilde{g}) = \sqrt{\sum_{p=1}^{n} R^2\,(\tilde{f}_p - \tilde{g}_p)} = \sqrt{\sum_{p=1}^{n} (R(\tilde{f}_p) - R(\tilde{g}_p))^2} \tag{23}$$

is called a fuzzy distance with relation to R where $\tilde{f} = (\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_n), \tilde{g} = (\tilde{g}_1, \tilde{g}_2, \dots, g_n) \in F^n(R)$.

It is clear that the description of d_R is a direct expansion of the formal Euclidean distance. Based on Lemma 2.2.1, the ranking mapping R to be described by (24) is linear. So, for ranking function R, the fuzzy distance can rewrite as follows:

$$R(\tilde{p}) = \alpha p_1 + 2\beta p_2 + \alpha p_3 \tag{24}$$

$$d_R^2([\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_n], [\tilde{g}_1, \tilde{g}_2, \dots, \tilde{g}_n]) = \sum_{p=1}^{n}[\alpha((f_p)_1 + (f_p)_3 - (g_p)_1 - (g_p)_3) + 2\beta((f_p)_2 + -(g_p)_2)]^2 \tag{25}$$

where $\tilde{f}_p = ((f_p)_1, (f_p)_2, (f_p)_3)$ and $\tilde{g}_p = ((g_p)_1, (g_p)_2, (g_p)_3)$ for p=1, 2, …, n. The (25) specifies the value $d_R^2(\tilde{f}, \tilde{g})$ clearly; but the compact form of definition 3.3.1 is applied.

The proposed UPS clustering algorithm solves

$$\text{Fitness function(UPSC)} = \frac{\sum_{L=1}^{c}[\sum_{\forall \tilde{d}_k \in C_{iL}} d_R^2(\tilde{d}_k, \tilde{m}_{iL})/|C_{iL}|]}{c} \tag{26}$$

$$\text{Fitness function(UPSC)} = \frac{\sum_{L=1}^{c}[\sum_{\forall \tilde{d}_k \in C_{iL}} \sum_{p=1}^{n}[\alpha((d_k^p)_1 + (d_k^p)_3 - (m_{iL}^p)_1 - (m_{iL}^p)_3) + 2\beta((d_k^p)_2 - (m_{iL}^p)_2)]^2/|C_{iL}|]}{c} \tag{27}$$

where in $\tilde{d}_k \in F^{n\times N}(R)$.

Now, the proposed algorithm (UPSC) for clustering triangular fuzzy data is described.

It is clear that any real number can be written as triangular fuzzy number, so Algorithm 1 is converted to the standard PSO algorithm for real data.

Algorithm 1: Uncertain particle swarm clustering (UPSC) algorithm

Choose the number of clusters c and set iteration = 1
Initialize the cluster centroids of each particle randomly
For iteration =1 to iteration $_{max}$ do
Begin
For each particle i do
Begin
For each data vector $\tilde{d}_k$ do
Begin
Calculate the $d_R^2(\tilde{d}_k, \tilde{m}_{iL})$ to all cluster centroids $\tilde{m}_{iL}$ using fuzzy distance (28) obtained using ranking functions.

$$d_R^2(\tilde{d}_k, \tilde{m}_{iL}) = \sum_{p=1}^{n}[\alpha\left((d_k^p)_1 + (d_k^p)_3 - (m_{iL}^p)_1 - (m_{iL}^p)_3\right) + 2\beta((d_k^p)_2 - (m_{iL}^p)_2)]^2 \quad (28)$$

Assign $\tilde{d}_k$ to cluster C_{iL} such that:

$$d_R^2(\tilde{d}_k, \tilde{m}_{iL}) = \min\ \{d_R^2(\tilde{d}_k, \tilde{m}_{iV})\} \qquad \text{for } V = 1,2,\dots,c \quad (29)$$

End
End
Calculate the fitness function using (26)
Update the global best and local best positions
Update the cluster centroids using (10) and (11)
End

It is a preference for the proposed algorithm (UPSC) that if data is a specific type (such as real numbers, interval numbers, or TFNs), then prototypes are the same as type; prototypes are linear compounds of data.

Experimental Results

In this section, datasets, evaluation criteria, and numerical results are explained.

F. Datasets

To evaluate the proposed method (UPSC), several datasets, including six artificial datasets and nine UCI datasets, are used. These datasets have been converted to triangular fuzzy numbers and then used.

Nine Benchmark Datasets

These datasets are iris, wine, vehicle, yeast, image, abalone, libras, glass, ecoli. These datasets are existent at the UCI machine learning repository [45]. The specifications of these datasets are existent in Table 1.

Table 1: Characteristics of fuzzy benchmark datasets

Dataset	#Objects	#Classes	#Attributes
Iris	150	4	3
Wine	178	13	3
Vehicle	846	18	4
Yeast	1484	8	10
Image	2310	19	7
Abalone	4124	7	17
Libras	360	90	15
Ecoli	327	7	5
Glass	214	9	6

Six Artificial Datasets

For three artificial datasets using MATLAB functions, a uniformly distributed pseudorandom integer has been created. The first dataset has 20 points distributed among two clusters Fig. 2. The second dataset has 500 points distributed among five clusters Fig. 3, and the third dataset has 150 points distributed among three clusters Fig. 4.

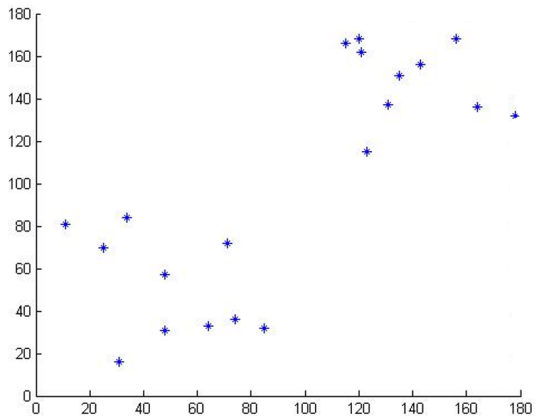


Fig. 2: Seed artificial dataset 1, which created using the uniform distribution of MATLAB software.

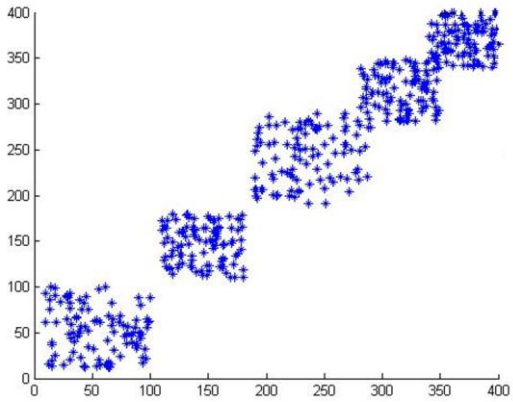


Fig. 3: Seed artificial dataset 2, which created using the uniform distribution of MATLAB software.

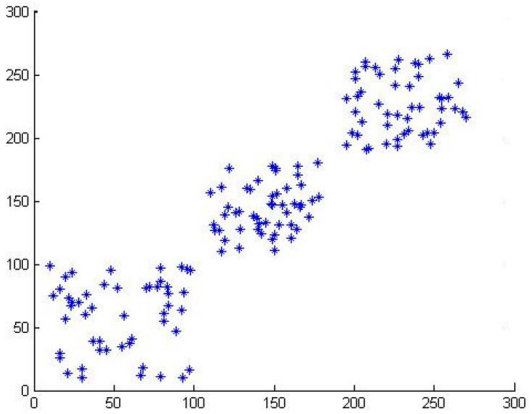


Fig. 4: Seed artificial dataset 3, which created using the uniform distribution of MATLAB software.

For the other two artificial datasets, the same data point formations introduced in [14] are considered. At first, two standard quantitative datasets in R^2 be regarded. Each dataset includes 350 points distributed among three clusters of unequal sizes and shapes: one cluster with size 50 and a spherical shape and two clusters with sizes 150 and ellipsis shapes. Data points of each cluster in datasets 4 and 5 were drawn following a bi-variate normal distribution of autonomous parts with th following mean vector and covariance matrix:

$$\mu = \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} \quad \text{and} \quad \sum_{11} = \begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix} \qquad (30$$

Artificial dataset 4 exhibits well-separated clusters Fig. 5.

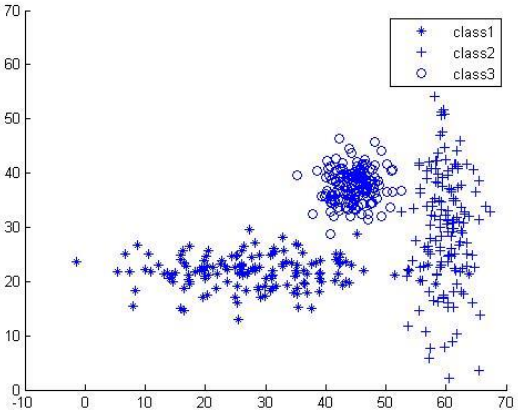


Fig. 5: Seed artificial dataset 4, which exhibit well-separated classes.

Data points of each cluster in dataset 4 were drawn following the subsequent parameters:

$$\text{Class 1: } \mu_1 = 28\,, \mu_2 = 22, \sigma_1^2 = 100, \sigma_2^2 = 9 \qquad (31)$$

$$\text{Class 2: } \mu_1 = 60\,, \mu_2 = 30, \sigma_1^2 = 9, \sigma_2^2 = 144 \qquad (32)$$

$$\text{Class 3: } \mu_1 = 45\,, \mu_2 = 38, \sigma_1^2 = 9, \sigma_2^2 = 9 \qquad (33)$$

Artificial dataset 5 exhibits overlapping clusters Fig. 6.

Data points of each cluster in dataset 5 were drawn following the subsequent parameters:

$$\text{Class 1: } \mu_1 = 45\,, \mu_2 = 22, \sigma_1^2 = 100, \sigma_2^2 = 9 \qquad (34)$$

$$\text{Class 2: } \mu_1 = 60\,, \mu_2 = 30, \sigma_1^2 = 9, \sigma_2^2 = 144 \qquad (35)$$

$$\text{Class 3: } \mu_1 = 52\,, \mu_2 = 38, \sigma_1^2 = 9, \sigma_2^2 = 9 \qquad (36)$$

The sixth dataset has non-linearly separable clusters. This dataset is very suitable for density-based clustering algorithms like DBSCAN. This dataset has 500 points and two clusters Fig. 7.

To create artificial data in this paper, an attempt has been made to use different types of data with different numbers, different distributions, different dispersions, different complexities, and different amounts of interference.

For example, dataset 1 is a straightforward artificial dataset. The purpose of creating it was merely to demonstrate the ability of the proposed method for solving the desired problem. In datasets 2 and 3, the number of samples and clusters has increased compared to dataset 1. In dataset 5, more interference was considered for data. In dataset 6, data distribution was considered more difficult than in other datasets.

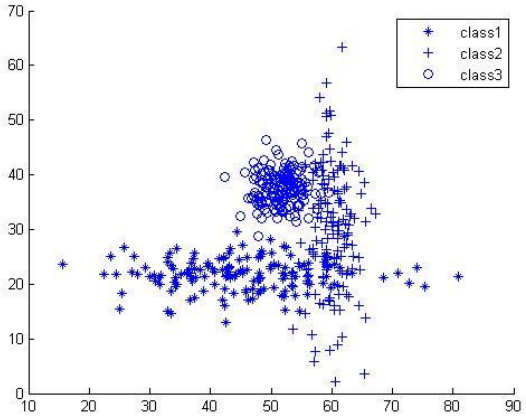


Fig. 6: Seed artificial dataset 5, which has overlapping classes.

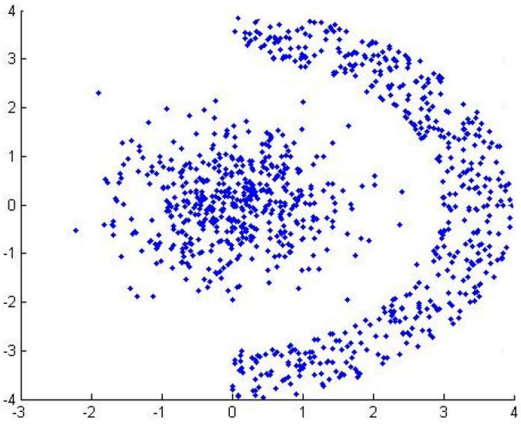


Fig. 7: Seed artificial dataset 6.

G. Converting Real Numbers to Triangular Fuzzy Number

A simple method is used to convert all datasets to triangular fuzzy datasets. Triangular fuzzy number in LR format is show in Fig. 8. In LR format we have:

$$P^* = P_3 - P_2\,, P_* = P_2 - P_1, P_0 = P_2 \qquad (37)$$

If P_{0_i} be a crisp or certain data, firstly, the maximum and minimum values of each crisp dataset $(\min(P_0) \text{ and } \max(P_0))$ are obtained using equations (38, 39) (N is the number of data sets). The P_* and P^* parameters are set to a random value using (40) and uniform distribution (41). The "rand" function of MATLAB software is used in (41) to generate random real numbers with a uniform distribution.

$$\min(P_0) = \min_{i=1,2,\ldots,N} P_{0_i} \qquad (38)$$

$$\max(P_0) = \max_{i=1,2,\dots,N} P_{0i} \quad (39)$$

$$\text{TFN range} = \text{alpha} \times (\max(P_0) - \min(P_0)) \quad (40)$$

$$P_* = P^* = \text{rand} \times \text{TFN range} \quad (41)$$

Now crisp data is converted to triangular fuzzy data (TFN). Triangular fuzzy data ($\widetilde{P}$) exhibits in the following format (42).

$$\widetilde{P} = (P_*, P_0, P^*)_{LR} = (P_2 - P_1, P_2, P_3 - P_2) \quad (42)$$

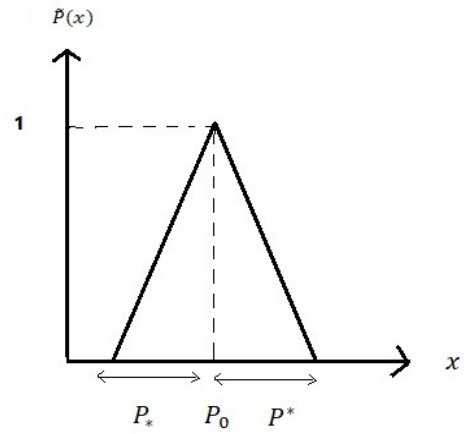


Fig. 8: Triangular fuzzy number in LR format ($\widetilde{P} = (P_0, P_*, P^*)_{LR}$).

Before using data to test the proposed method (UPSC), normalization preprocessing is performed on it.

H. Evaluation criteria

In this paper, the accuracy criterion mentioned in [16], which is one of the most popularly used criteria, is utilized to measure the accuracy of the clustering results. To calculate accuracy, each cluster is assigned to the class which is most repeated in the cluster. Then the accuracy of this assignment is evaluated by counting the number of correctly assigned objects and dividing them on all objects n. formally

$$\text{accuracy}(M, CL) = \frac{1}{n} \sum_c \max_k |m_c \cap cl_k| \quad (43)$$

where $M = \{m_1, m_2, \dots, m_c\}$ is the set of clusters and CL=$\{cl_1, cl_2, \dots, cl_k\}$ is the set of classes.

The Rand Index (RI) criterion is another criterion computed to exhibit the resemblance between the two procedures of labeling (44). The rand index for error-free clustering is equal to 1.

$$\text{Rand Index}(M, CL) = \frac{a+b}{n(n-1)/2} \quad (44)$$

In (44), a is the number of pairs that are together in both classes and clusters, and b is the number of pairs that are separated from each other both in classes and clusters.

I. Numerical Results

In this section, the proposed method and other methods are evaluated based on the mentioned criteria.

In all simulations, c_1=c_2=2 and w=0.7 for UPSC, and as discussed in section 2, α=β=1/4 for ranking function is set. The experiment is repeated ten times to decrease the variation from trial to trial. Table 2 shows the accuracy results on six artificial datasets for the proposed method (UPSC), UK-means, and UK-medoids. The Accuracy or purity is one of the most popular criteria for evaluating clustering methods. Two sample results for each dataset in Table 2 are reported.

These results are related to two different alpha parameters in (40). The first row is related to $\text{alpha} = 0.01$, and the second row is associated with $\text{alpha} = 0.03$.

The higher the alpha, the higher the margin of fuzzy triangular numbers (P_* and P^* in Fig. 8). The average, best and worst answers for ten times implementation of these algorithms are reported in Table 2. The mean is the average of the answers received from the ten times the implementation of these algorithms. Also, best and worst show the best and worst answers received from the ten times of implementing these algorithms, respectively. To evaluate the proposed method (UPSC), the Uk-Means and Uk-Medoids algorithms also implemented. In fuzzy artificial dataset 1, which is straightforward, all three methods performed the clustering without error. In artificial dataset 2, when the alpha coefficient increases to 0.03 (data uncertainty increases), firstly, our proposed method can find the answer without error in this case. Secondly, the mean, the best and the worst answer are slightly different together and can be said that the stability of our proposed method is higher. In the artificial dataset 3, as can be seen from Table 2, for alpha=0.01, answers of all three methods are without error. Still, by increasing the alpha coefficient to 0.03, the superiority of UPSC is quite apparent, because in this case, mean, best and worst answer is the same.

This means complete stability of the proposed method in the case of dataset 3, While in the UK-Means and UK-medoids methods, the best and worst answers are different. In artificial dataset 4, although the UK-Means and UK-medoids methods have better stability, the mean and the best answer obtained using UPSC are better. For the artificial dataset 5, which has high interference, and the artificial dataset 6, which has a more difficult distribution, according to Table 2, the UPSC method has the better answers and good stability.

After artificial datasets, to have a more realistic valuation, the UPSC method is applied to the UCI datasets as mentioned above. The optimal cluster prototypes are available for these datasets. Table 3 shows the accuracy results on nine fuzzy benchmark datasets.

The results are achieved by repeating the experiments 10 times.

Table 2: Clustering accuracy results on fuzzy artificial datasets using UPSC, UK-Means, and UK-medoids methods

Dataset		UPSC			UK-Means			UK-medoids		
		Mean	best	worst	Mean	best	worst	Mean	best	worst
Artificial dataset 1	alpha=0.01	1	1	1	1	1	1	1	1	1
	alpha=0.03	1	1	1	1	1	1	1	1	1
Artificial dataset 2	alpha=0.01	0.9712	0.9920	0.7840	0.8780	0.9960	0.7960	0.8688	0.9960	0.7680
	alpha=0.03	0.9976	1	0.9960	0.8748	1	0.7720	0.9228	0.9800	0.7960
Artificial dataset 3	alpha=0.01	1	1	1	1	1	1	1	1	1
	alpha=0.03	1	1	1	0.9547	0.9867	0.6667	0.9667	1	0.6667
Artificial dataset 4	alpha=0.01	0.8386	0.8786	0.8214	0.8086	0.8143	0.8000	0.8057	0.8143	0.8000
	alpha=0.03	0.8993	0.9071	0.8286	0.8621	0.8786	0.8571	0.8514	0.8571	0.8286
Artificial dataset 5	alpha=0.01	0.7843	0.7857	0.7714	0.7300	0.7500	0.7000	0.7357	0.7571	0.7143
	alpha=0.03	0.7814	0.7857	0.7786	0.7093	0.7143	0.7071	0.7071	0.7643	0.6929
Artificial dataset 6	alpha=0.01	0.8480	0.8640	0.8410	0.8377	0.8450	0.8290	0.7758	0.8590	0.5420
	alpha=0.03	0.8560	0.8680	0.8430	0.8442	0.8540	0.8400	0.8372	0.8600	0.8170

As can be seen in Table 3, UPSC has higher accuracy than the other two methods for iris data. Besides, the best and worst answers of UPSC for iris data are less different than the other two methods. In the case of wine data, UPSC method is more accurate and stable than the UK-medoids method and more accurate than the UK-Means Method. In the vehicle dataset, when the data uncertainty increases with increasing alpha, the ability of our proposed method increases, and we reach a higher accuracy than the other two methods. In the yeast dataset, the best answer with a slight difference is related to UPSC method, but the mean of our method is better with more differences than the other two methods. For the image dataset, UPSC method has the highest accuracy. In the abalone dataset, the UPSC method has a significantly better answer, and with increasing the alpha coefficient, the accuracy of the UPSC method increases. In contrast, for the UK-Means and the UK-medoids methods, the accuracy decreases with increasing alpha, and this demonstrates the ability of the UPSC method for clustering fuzzy data.

In the glass and ecoli datasets, according to Table 3, the proposed method has a better answer, and like the other the dataset, the clustering accuracy increases with increasing alpha. Only in the libras dataset the accuracy of UPSC method less than that of the other two methods. Although the UK-Means Method has the highest accuracy for the libras dataset, in this method and the UK-medoids method, accuracy decreased with increasing alpha. While with increasing alpha coefficient, UPSC's ability in clustering increased.

In the continuation of this section, Table 4 and Table 5 are presented to facilitate the comparison of UPSC method with other methods.

The best answer for the proposed method was for alpha=3. Therefore, Table 4 and Table 5 are reported for alpha=3. Table 4 is a summary of Table 2. The achieved results for UPSC in challenging with the other methods show the capability and robustness of UPSC method. As Table 4 shows, in all artificial datasets, UPSC method gives the best answer. Table 5 compares the results of UPSC method on the benchmark dataset with the UK-means, the UK-medoids, and uncertain ACO (Ant Colony Optimization) methods. In Table 5, the best accuracy, the mean accuracy and RI index are reported. The propinquity of the best accuracy and the mean accuracy in Table 5 illustrate the higher stability of UPSC method.

According to Table 5 in eight datasets, UPSC method has higher accuracy. Although in libras dataset the UK-means method has the highest accuracy, but with increasing the alpha coefficient, the ability of UPSC method for clustering libras data increase. In contrast, for the UK-means method, with increasing alpha, the clustering accuracy decrease. In Table 5, whatever the rand index (RI) is close to 1, The clustering error is less.

According to the calculations reported in Table 5, the proposed clustering method has a better Rand index in datasets. Due to the random and search-based nature of the proposed method, the volume of calculations and the execution time of the algorithm increase. Of course, since the calculations are done offline, this issue becomes less important. In other words, this time is related to clustering the existing data and finding the optimal centers of the clusters. After the clustering is done, the clustering of unknown and new samples is done in very little time.

Table 3: Clustering accuracy results on fuzzy benchmark datasets using UPSC, UK-Means, and UK-medoids methods

Dataset		UPSC			UK-Means			UK-medoids		
		Mean	best	worst	Mean	best	worst	Mean	best	worst
Iris	alpha=0.01	0.8973	0.9200	0.8867	0.8647	0.8867	0.6667	0.8300	0.9067	0.6667
	alpha=0.03	0.8960	0.9267	0.8800	0.8867	0.8867	0.8867	0.8493	0.9000	0.6667
Wine	alpha=0.01	0.9534	0.9663	0.9438	0.9449	0.9494	0.9326	0.8326	0.9270	0.6011
	alpha=0.03	0.9579	0.9663	0.9494	0.9478	0.9551	0.9326	0.8629	0.9663	0.6180
Vehicle	alpha=0.01	0.3827	0.4031	0.3700	0.3959	0.4078	0.3652	0.3858	0.4102	0.3641
	alpha=0.03	0.3987	0.4232	0.3712	0.3858	0.4043	0.3641	0.3892	0.4078	0.3629
Yeast	alpha=0.01	0.4739	0.5168	0.4387	0.4678	0.5004	0.4135	0.4725	0.5068	0.4422
	alpha=0.03	0.5054	0.5263	0.4650	0.4935	0.5218	0.4401	0.4980	0.5200	0.4161
Image	alpha=0.01	0.6541	0.7039	0.6117	0.6214	0.6814	0.5745	0.6403	0.6926	0.5667
	alpha=0.03	0.6518	0.7061	0.6160	0.6206	0.6771	0.5745	0.6292	0.6948	0.5502
Abalone	alpha=0.01	0.2602	0.2679	0.2512	0.2164	0.2204	0.1998	0.2225	0.2345	0.2072
	alpha=0.03	0.2611	0.2701	0.2500	0.2142	0.2200	0.202	0.2220	0.2330	0.2167
Libras	alpha=0.01	0.4158	0.4417	0.3917	0.4636	0.5083	0.4306	0.4414	0.4833	0.4028
	alpha=0.03	0.4156	0.4667	0.3611	0.4758	0.5028	0.4500	0.4342	0.4861	0.3861
Glass	alpha=0.01	0.5308	0.5561	0.5000	0.5108	0.5501	0.4800	0.5290	0.5561	0.5093
	alpha=0.03	0.5407	0.5794	0.4953	0.5192	0.5514	0.5000	0.5286	0.5415	0.5040
Ecoli	alpha=0.01	0.7908	0.8318	0.7523	0.7801	0.8043	0.7523	0.7839	0.8218	0.7454
	alpha=0.03	0.7920	0.8349	0.7584	0.7755	0.7890	0.7645	0.7801	0.8226	0.6453

Table 4: Comparing The best accuracy of UPSC method with the best accuracy of other methods for fuzzy artificial datasets

Dataset	UPSC (proposed method)	UK-Means	UK-medoids
Artificial dataset 1	**1**	**1**	**1**
Artificial dataset 2	**1**	**1**	0.98
Artificial dataset 3	**1**	0.9867	**1**
Artificial dataset 4	**0.9071**	0.8786	0.8571
Artificial dataset 5	**0.7857**	0.7143	0.7643
Artificial dataset 6	**0.8680**	0.8540	0.8600

Fig. 9 shows the implementation time of the proposed clustering method (UPSC) for all data. All tests and simulations in this paper (for proposed method and other conventional method) by the computer with Core (TM) 2 Duo 2.20GHz central processing unit and 4GB memory in MATLAB software environment performed. As shown in Fig. 9, the larger the dataset volume according to Table 1, the more time is required to implement the UPSC method.

For example, according to Table 1 abalone dataset contains 4124 objects.

This dataset has 7 features, and the goal is to separate the abalone dataset into 17 clusters. Therefore, more time is needed to run the UPSC method than (for example) the iris dataset; Because the iris dataset has 150 4-dimensional samples that should be divided into 3 clusters.

Artificial datasets are also small in volume, so little time is required to implement the UPSC method.

Due to the large number of datasets used in this paper, only clustering output diagrams for artificial datasets 1, 2 and, 3 are shown.

Fig. 10, Fig. 11, and Fig. 12 show the clustering diagrams for these datasets. The diagrams in Fig. 10, Fig. 11, and Fig. 12 associated with the UK-means and the UK-medoids methods are for alpha=0.03.

Table 5: Comparing the UPSC method with other methods for fuzzy benchmark datasets

	UPSC (proposed method)			UK-means			UK-medoids			Uncertain ACO		
	Best accuracy	Mean accuracy	RI	Best accuracy	Mean accuracy	RI	Best accuracy	Mean accuracy	RI	Best accuracy	Mean accuracy	RI
Iris	**0.9267**	**0.8960**	**0.9400**	0.8867	0.8867	0.9273	0.9000	0.8493	0.9384	0.9211	0.8945	0.9388
Wine	**0.9663**	**0.9579**	**0.9478**	0.9494	0.9449	0.9344	0.9270	0.8326	0.9412	0.9479	0.9389	0.9320
Vehicle	**0.4232**	**0.3987**	0.3793	0.4043	0.3858	**0.3794**	0.4078	0.3892	0.3755	0.3992	0.3850	0.3782
yeast	**0.5263**	**0.5054**	**0.6548**	0.5218	0.4935	0.6117	0.5200	0.4980	0.6188	0.5243	0.5030	0.6517
image	**0.7061**	**0.6518**	**0.6784**	0.6771	0.6206	0.6617	0.6948	0.6292	0.6702	0.6954	0.6615	0.6720
Abalone	**0.2701**	**0.2611**	**0.7896**	0.2200	0.2142	0.7869	0.2330	0.2220	0.7725	0.2345	0.2301	0.7851
Libras	0.4667	0.4156	0.5873	**0.5028**	**0.4758**	**0.5990**	0.4861	0.4342	0.5897	0.4634	0.4020	0.5855
Glass	**0.5794**	**0.5407**	**0.5879**	0.5514	0.5192	0.5791	0.5415	0.5286	0.5788	0.5702	0.5326	0.5833
Ecoli	**0.8349**	**0.7920**	**0.8343**	0.7890	**0.7755**	0.8143	0.8226	0.7801	0.8275	0.8240	0.7899	0.8225

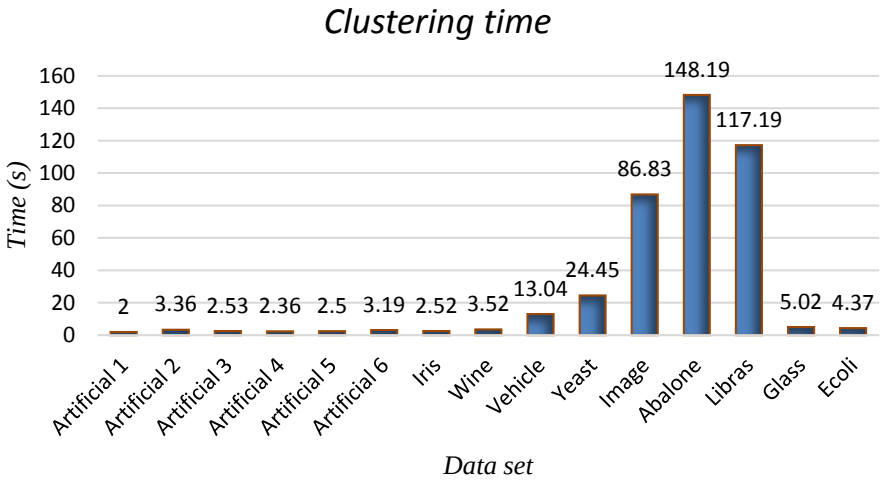


Fig. 9: Clustering time for all fuzzy datasets in UPSC method.

To see the results of the proposed method (UPSC) for alpha=0.01, we have also shown it in Fig. 10, Fig. 11, and Fig. 12. These clustering diagrams show that UPSC method can find centroids of clusters related to these datasets without error or minimum error.

Dataset 1 is a straightforward artificial dataset. The purpose of creating this dataset was mere to demonstrate the ability of UPSC method in fuzzy/uncertain data clustering. In fact, in clustering, the results are the centers of the clusters. The two obtained cluster centers for the artificial dataset 1 are shown in Fig. 10 with two squares, blue and red. As you can see in Fig. 10, UPSC method with two selected alpha parameters 0.01, and 0.03 (top row in Fig. 10), has been able to perform the clustering operation like the UK-means and UK-medoids methods (bottom row in Fig. 10) correctly and achieved the best centers of clusters.

All blue data is assigned to the cluster with the blue center and, all red data is transferred to the cluster with a red center.

Fig. 11 shows the clustering diagram related to artificial dataset 2. Fig. 11 shows the superiority of UPSC method clearly. In the above row of Fig. 11, which is related to UPSC method with alpha 0.01 and 0.03, all the color data are assigned to their respective clusters. The bottom row of Fig. 11, is corresponding to the UK-means and UK-medoids methods that have errors in this implementation and have not been able to obtain the centers of the clusters correctly. The five correct clusters created in the artificial dataset 2 are shown in Fig. 3. The last diagram that shown as an example is related to the artificial dataset 4. By comparing Fig. 12 with Fig. 5, the superiority of UPSC method is identified. The top row of Fig. 12 (UPSC method) has performed clustering with fewer errors than the bottom row methods (UK-means and UK-medoids).

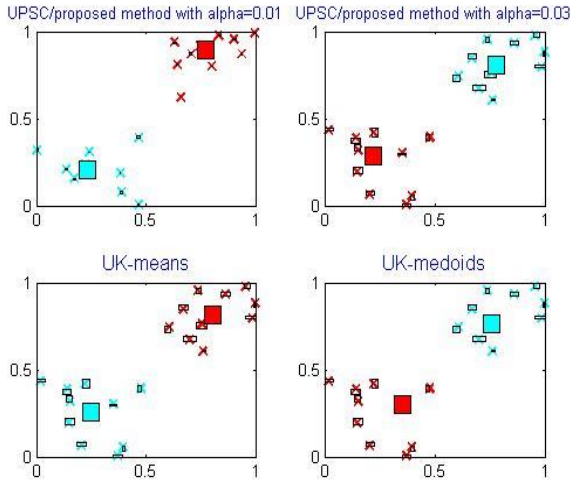


Fig. 10: Clustering diagrams for fuzzy artificial dataset 1.

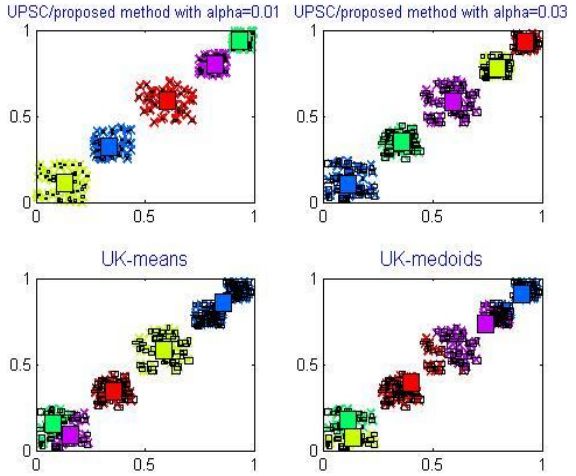


Fig. 11: Clustering diagrams for fuzzy artificial dataset 2.

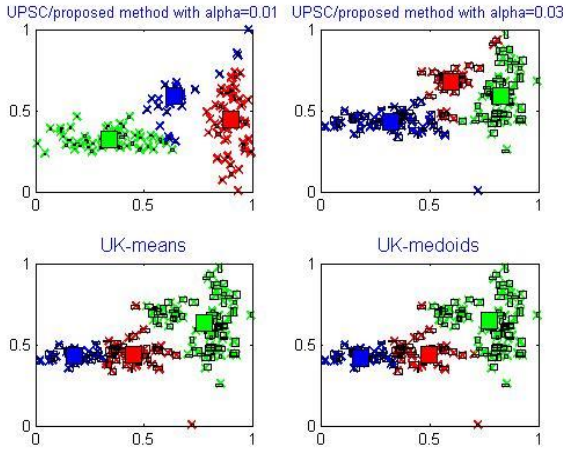


Fig. 12: Clustering diagrams for fuzzy artificial dataset 4.

Conclusion

In this paper, for the first time, the particle swarm optimization algorithm is upgraded for clustering of uncertain/fuzzy datasets. For this purpose, a new cost function is defined for the traditional particle swarm clustering method. The proposed cost function can find the optimal centers of the clusters for data whose features are expressed in triangular fuzzy numbers (TFNs). On the other hand, an uncertain particle swarm clustering (UPSC) method is proposed, which can cluster fuzzy datasets. The proposed method (UPSC) can be also used for clustering uncertain datasets with interval data, trapezoidal fuzzy data, or real data. The achieved results show that the proposed method (UPSC) is superior to the challenging and standard methods such as UK-means, UK-medoids, and Uncertain ACO algorithms. So, the UPSC is an efficient solution for the problem of clustering uncertain data. For example, in iris data, the clustering accuracy has increased by 2.67% compared to the UK-means method. In the case of wine data, the accuracy increase with the proposed method is 1.69%. As another example, it can be said that the increase in accuracy for abalone data was 4%. Comparing the results with the Rand index also shows the superiority of the proposed method. The most important limitations of heuristic algorithms, including the PSO algorithm, are capturing local optimum and matter of being offline. Of course, it should be mentioned that the most of the clustering algorithms are offline and can meet the needs of the problem. In this paper, the PSO algorithm is upgraded for clustering uncertain/fuzzy data. One of the works that can be done in the future is using other versions of the PSO algorithm that are more complex and may give better results. Using soft computing methods (e.g., fuzzified PSO / fuzzy and PSO) is another works that can be done in the future. Also, the proposed method is general and modular. Therefore, other heuristic methods such as inclined planes system optimization (IPO) or gravitational search algorithm (GSA) with minimum changes can be used instead of PSO to solve uncertain/fuzzy data clustering.

Author Contributions

Dr. Zahiri and Dr. Shahraki were the supervisor and adviser of the current research paper. They sketched the research framework and the roadmap. Also, they analyzed the results. N. Ghanbari searched in authentic journals to gather all relevant papers. Also, she collected the data and wrote the manuscript. Dr. Zahiri, Dr. Shahraki, and N. Ghanbari interpreted the results.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent,

misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

PSO	Particle Swarm Optimization
UPSC	Uncertain Particle Swarm Clustering
TFNs	Triangular Fuzzy Numbers
UK- means	Uncertain K-Means
UK-Medoids	Uncertain K-Medoids
FCM	Fuzzy Clustering Method
IPO	Inclined Planes System Optimization
GSA	Gravitational Search Algorithm
RI	Rand Index
ACO	Ant Colony Optimization

References

[1] A. Dutt, M. A. Ismail, T. Herawan, "A systematic review on educational data mining," IEEE Access, 5: 15991-16005, 2017.

[2] Z. Wang "Determining the clustering centers by slope difference distribution," IEEE Access, 5: 10995-11002, 2017.

[3] T. T. Zhang, B. Yuan, "Density-based multiscale analysis for clustering in strong noise settings with varying densities," IEEE Access, 6: 25861-25873, 2018.

[4] P. N. Tan, M. Steinbach, V. Kumar, "Introduction to datamining," Addison-Wesley, 2005.

[5] H. Shahraki, S. H. Zahiri, "Classification of trapezoidal fuzzy data based on heuristic classifiers," Kasmera, 43(1): 128-144, 2015.

[6] F. Gullo, "An information-theoretic approach to hierarchical clustering of uncertain data," Inf. Sci., 402: 199-215, 2017.

[7] Y. Mao, Y. Liu, M.A. Khan, J. Wang, D. Mao, J. Hu, "Uncertain interval data EFCM-ID clustering algorithm based on machine learning," J. Rob. Mechatron., 31(2): 339-347, 2019.

[8] L. Yue, L. Zitu, L. Shuang, G. Yike, L. Qun, W. Guoyin, "Cloud-Cluster: An uncertainty clustering algorithm based on cloud model," Knowledge-Based Syst., 263, 2023.

[9] G. S. Nijaguna, K. Thippeswamy, "Multiple kernel fuzzy clustering for uncertain data classification," Int. J. Comput. Eng. Technol. (IJCET), 10(01): 253-261, 2019.

[10] C. Ko, J. Baek, B. Tavakkol, Y. S. Jeong, "Cluster Validity Index for Uncertain Data Based on a Probabilistic Distance Measure in Feature Space," Sensors, 23(7): 3708, 2023.

[11] B. Tavakkol, Y. Son, "Fuzzy kernel K-medoids clustering algorithm for uncertain data objects," Pattern Anal. Appl., 24(3): 1287-1302, 2021.

[12] J. Zhou, L. Chen, C. L. Philip Chen, Y. Wang, H. X. Li, "Uncertain data clustering in distributed peer-to-peer networks," IEEE Trans. Neural Networks Learn.Syst., 29(6): 2392-2406, 2018.

[13] H. Shahraki, S. H. Zahiri, "Fuzzy decision function estimation using fuzzified particle swarm optimization," Int. J. Mach. Learn. Cybern., 8: 1827-1838, 2017.

[14] R. M.C.R. de Souza, F. de A.T. de Carvalho, "Clustering of interval data based on city–block distances," Pattern Recognit. Lett., 25: 353–365, 2004.

[15] F. de A.T. de Carvalho, R. M.C.R. de Souza, M. Chavent, Y. Lechevallier, "Adaptive hausdorff distances and dynamic clustering of symbolic interval data," Pattern Recognit. Lett., 27: 167–179, 2006.

[16] X. Zhang, H. Liu, X. Zhang, "Novel density-based and hierarchical density-based clustering algorithms for uncertain data," Neural Networks, 93: 240–255, 2017.

[17] J. Tayyebi, E. Hosseinzadeh, "A fuzzy c-means algorithm for clustering fuzzy data and its application in clustering incomplete data," J. AI Data min., 8(4): 515-523, 2021.

[18] R. Adrian, S. Sulistyo, I.W. Mustika, S. Alam, "ABNC: Adaptive border node clustering using genes fusion based on genetic algorithm to support the stability of cluster in VANET," Int. J. Intell. Eng. Syst., 13(1): 354-363, 2020.

[19] T. P. Q. Nguyen, R. J. Kuo, "Partition-and-merge based fuzzy genetic clustering algorithm for categorical data," Appl. Soft Comput. J., 75: 254–264, 2019.

[20] I. Behravan, S. H. Zahiri, S. M. Razavi, R. Trasarti, "Finding roles of players in football using automatic particle swarm optimization-clustering algorithm," Big Data, 7(1): 35-56, 2019.

[21] Z. Liu, B. Xiang, Y. Song, H. Lu, Q. Liu, "An improved unsupervised image segmentation method based on multi-objective particle swarm optimization clustering algorithm," CMC-Comput. Mater. Continua, 58(2): 451-461, 2019.

[22] B. Anari, J. Akbari torkestani, A. M. Rahmani, "A learning automata-based clustering algorithm using ant swarm intelligence," Expert Syst., 35(6): e12310, 2018.

[23] M. S. Tomar, P. K. Shukla, "Energy efficient gravitational search algorithm and fuzzy based clustering with hop count-based routing for wireless sensor network," Multimedia Tools Appl., 78: 27849–27870, 2019.

[24] H. Mittal, M. Saraswat, "An automatic nuclei segmentation method using intelligent gravitational search algorithm based super pixel clustering," Swarm Evol. Comput., 18(9): S2210-6502, 2018.

[25] J. Kennedy, R. C. Eberhart, "Particle swarm optimization," in Proc. IEEE Internal Conference on Neural Networks, 4: 1942-1948, 1995.

[26] W. Xiong, "Initial clustering based on the swarm intelligence algorithm for computing a data density parameter," Comput. Intell. Neurosci.: 1-8, 2022.

[27] D. W. Van der Merwe, A. P. Engelbrecht, "Data clustering using particle swarm optimization in evolutionary computation," in Proc. The 2003 Congress on Evolutionary Computation, 2003. CEC '03, 2003.

[28] R. E. Bellman, L. A. Zadeh, "Decision making in a fuzzy environment," Manag. Sci., 17: 141-164, 1970.

[29] H. Tanaka, H. Ichihashi, "A formulation of fuzzy linear programming problem based on comparison of fuzzy numbers," Control Cyber., 13: 185-194, 1984.

[30] Y.J. Lai, C.L. Hwang, "Fuzzy mathematical programming methods and applications," Springer, Berlin, 1992.

[31] S. C. Fang, C. F. Hu, H. F. Wang, S. Y. Wu, "Linear programming with fuzzy coefficients in constraints," Comput. Math. Appl., 37(10): 63-76, 1999.

[32] T. Shaocheng, "Interval number and fuzzy number linear programming," Fuzzy Sets Syst., 66(3): 301-306, 1994.

[33] C. Garcia-Aguado, J. L. Verdegay, "On the sensitivity of membership functions for fuzzy linear programming problems," Fuzzy Sets Syst., 56(1): 47-49, 1993.

[34] H. R. Maleki, "Ranking functions and their applications to fuzzy linear programming," Far East J. Math. Sci, 4(3): 283-301, 2002.

[35] Y. L. P. Thorani, P. Phani Bushan Rao, N. Ravi Shankar, "Ordering generalized trapezoidal fuzzy numbers," Int. J. Contemp. Math. Scie., 7(12): 555 – 573, 2012.

[36] M. J. Ebadi, M. Suleiman, F. B. Ismail, A. Ahmadian, M. R. Baluch Shahryari, S. Salahshour, "A new distance measure for trapezoidal fuzzy numbers," Math. Probl. Eng., Article ID: 424186, 2013.

[37] T. Allahviranloo, M. A. Jahantigh, S. Hajighasemi, "A new distance measure and ranking method for generalized trapezoidal fuzzy numbers," Math. Probl. Eng., Article ID: 623757, 2013.

[38] N. Mahdavi-Amiri, S. H. Nasseri, "Duality in fuzzy number linear programming by use of a certain linear ranking function," Appl. Math. Comput., 180: 206-216, 2006.

[39] P. K. De. Debaroti Das, "Ranking of trapezoidal intuitionistic fuzzy numbers," in Proc. 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012.

[40] T. Hasuike, "Technical and cost efficiencies with ranking function in fuzzy data envelopment analysis," in Proc. Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011.

[41] D. Ponnialagan, J. Selvaraj, L. G. N. Velu, "A complete ranking of trapezoidal fuzzy numbers and its applications to multi-criteria decision making," Neural Comput. Appl., 30: 3303-3315, 2018.

[42] R. R. Yager, "A procedure for ordering fuzzy sets of the unit interval," Inf. Sci., 24: 143-161, 1981.

[43] H. Shahraki, S. H. Zahiri, "Design and simulation of an RF MEMS switch for removing the self: actuation and latching phenomena using PSO method," Iran J. Electr. Comput. Eng., 12: 56-63, 2013.

[44] D. W. van der Merwe, A. P. Engelbrecht, "Data clustering using particle swarm optimization," in Proc. The 2003 Congress on Evolutionary Computation, 2003. CEC '03: 215-220, 2003.

[45] S. Hettich, C. L. Blake, C. J. Merz, "UCI repository of machine learning database," Department of Information and Computer Science. University of California, Irvine, CA.

Biographies

Najme Ghanbari was born in Iran. She is the coach with the Department of Electronics Engineering, University of Zabol, Iran. She received the B.Sc. and M. Sc. degrees in electronics engineering from the University of sistan and baluchestan, Zahedan, Iran and Birjand University, Birjand, Iran, in 2002 and 2008, respectively. She is currently a Ph.D. student at Birjand University to receive a Ph.D. degree in electronics engineering. Her research interests include Pattern Recognition, Evolutionary Algorithms, and Swarm Intelligence Algorithms.

- Email: najme.ghanbari@birjand.ac.ir
- ORCID: 0000-0002-2804-1790
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Seyed Hamid Zahiri received the B.Sc., M.Sc. and Ph.D. degrees in Electronics Engineering from Sharif University of Technology, Tehran, Tarbiat Modarres University, Tehran, and Mashhad Ferdowsi University, Mashhad, Iran, in 1993, 1995, and 2005, respectively. Currently, he is a Professor with the Department of Electronics Engineering, University of Birjand, Birjand, Iran. His research interests include pattern recognition, evolutionary algorithms, swarm intelligence algorithms, and soft computing.

- Email: hzahiri@birjand.ac.ir
- ORCID: 0000-0002-1280-8133
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Hadi Shahraki received the B.Sc. degree in Electronics Engineering from University of Sistan and Baluchestan, Zahedan, Iran, in 2009 and M.Sc. degree in Electronics Engineering from Ferdowsi University of Mashhad, Mashhad, Iran, in 2012, and Ph.D. degree in Electronics Engineering from University of Birjand, in 2016.He is currently is assistant professor at the University of Sistan and Baluchestan, Zahedan, Iran. His research interests include pattern recognition and swarm intelligence algorithms.

- Email: hadi_shahraki@eng.usb.ac.ir
- ORCID: 0000-0001-9234-3577
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
N. Ghanbari, S. H. Zahiri, H. Shahraki, "Clustering of Triangular Fuzzy Data Based on Heuristic Methods," J. Electr. Comput. Eng. Innovations, 12(1): 1-14, 2024.

DOI: 10.22061/jecei.2023.9641.645

URL: https://jecei.sru.ac.ir/article_1894.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Review paper

Service and Energy Management in Fog Computing: A Taxonomy Approaches, and Future Directions

S. M. Hashemi [1], A. Sahafi [2,*], A. M. Rahmani [3], M. Bohlouli [4,5,6]

[1]Department of Computer Engineering, Qeshm branch, Islamic Azad university, Qeshm, Iran.

[2]Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.

[3]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.

[4]Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences, Zanjan, Iran.

[5]Research and Innovation Department, Petanux GmbH, Bonn, Germany.

[6]Research Center for Basic Sciences and Modern Technologies (RBST), Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran.

Article Info

Article History:

Received 11 February 2023
Reviewed 24 March 2023
Revised 22 May 2023
Accepted 27 May 2023

Keywords:

Fog computing
Internet of Things (IoT)
Systematic literature Review (SLR)
Service management
Energy management

*Corresponding Author's Email Address: *sahafi@iau.ac.ir*

Abstract

Background and Objectives: Today, the increased number of Internet-connected smart devices require powerful computer processing servers such as cloud and fog and necessitate fulfilling requests and services more than ever before. The geographical distance of IoT devices to fog and cloud servers have turned issues such as delay and energy consumption into major challenges. However, fog computing technology has emerged as a promising technology in this field.

Methods: In this paper, service/energy management approaches are generally surveyed. Then, we explain our motivation for the systematic literature review procedure (SLR) and how to select the related works.

Results: This paper introduces four domains of service management and energy management, including Architecture, Resource Management, Scheduling management, and Service Management. Scheduling management **has** been used in 38% of the papers. Therefore, they have the highest service management and energy management. Also, Resource Management is the second domain that has been able to attract about 26% of the papers in service management and energy management.

Conclusion: About 81% of the fog computing papers simulated their approaches, and the others implemented their schemes using a testbed in the real environment. Furthermore, 30% of the papers presented an architecture or framework for their research, along with their research. In this systematic literature review, papers have been extracted from five valid databases, including IEEE Xplore, Wiley, Science Direct (Elsevier), Springer Link, and Taylor & Francis, from 2013 to 2022. We obtained 1596 papers related to the discussed subject. We filtered them and achieved 47 distinct studies. In the following, we analyze and discuss these studies; then we review the parameters of service quality in the papers, and ultimately, we present the benefits, drawbacks, and innovations of each study.

Introduction

The Internet of Things (IoT) [1], [2] advancement has resulted in the urge of having fast and robust processing power for devices in important applications. In addition, data processing in the centralized cloud cannot meet requirements in such an environment. There are characteristics such as health monitoring and emergency response that should consider data delay and the data volume transferred to the cloud [3], [4]. Cisco proposed a novel paradigm known as fog computing (FC). This is an extended version of cloud computing [5], providing services toward the network edge. FC [6], [7] is a model for the management of a virtual and distributed environment to provide network services and computations between the cloud data centers [4], [8] and sensors [9].

Similar to Cloud [5] Fog provides data, services of computing, storage, and software for end-users and customers [10]. However, the Fog can be better than the cloud according to the following reasons:

- Closer topological proximity to end-users
- Geographical and large-scale distribution
- Support for mobility

These features will solve many issues to a large extent. In this regard, fog computing is generalized toward the network edge [11], [12] to reduce delay and congestion on the network and resolves several challenges such as high delay, low capacity, and network defects [4] Fig. 1 shows the cloud, Fog, and IoT services.

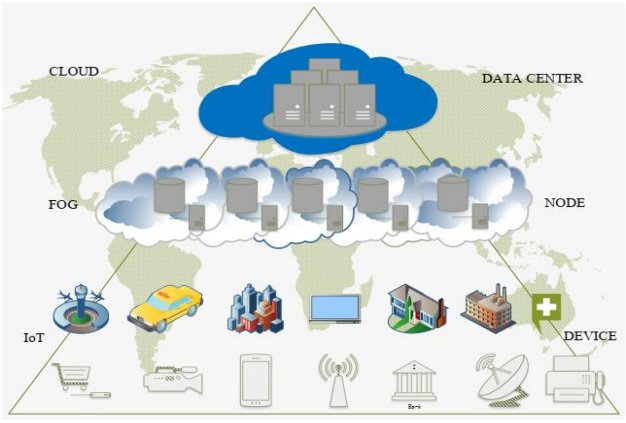


Fig. 1: Cloud, Fog and, IoT services.

There are several studies on specialized approaches/protocols. However, there is no review paper on service and energy management in Fog/edge Computing. The main purpose of this study is to discuss different service and energy management approaches in Fog/edge Computing. Some review studies usually discuss a separate topic, for example, energy approaches in [13] or service approaches/resource management in [14] and this topic has not been explored in a comprehensive study yet. In this SLR method, we investigate and discuss recent approaches/protocols in fog computing. Table 1 presents some studies in different fields of fog computing.

Table 1: Related investigations in fog computing service and energy management approach

Reference	Main Topic	Type	year of Publication	year Covered
[15]	Orchestration in FC	SLR	2022	**2015-2021**
[16]	simulation frameworks and research directions in FC	SLR	2021	**2015-2020**
[17]	Task scheduling approaches in FC	SLR	2020	**2012-2020**
[18]	Resource management approaches in FC	SLR	2019	**2014-2018**
[19]	Fundamental, network applications, and research challenges in FC	SURVEY	2018	**2014-2018**
[20]	Challenges in fog computing	SLR	2017	**2014-2017**
[21]	FC and its role in the IOT	SURVEY	2016	-

Related Studies

Researchers in [15] presented an SLR paper focusing on orchestration in fog computing. Their research period was from 2012 to 2021. They selected articles from 4 valid databases. By reviewing the literature of published articles, they limited their research to 50 articles written between 2015 and 2021. Their study consisted of 5 questions. Given the novelty of the topic of orchestration in fog computing, they presented a suitable conclusion for researchers, including the benefits, challenges, and future tasks.

Researchers in [16] used a systematic method to collect and identify articles on fog computing and simulation tools. Their study period was from 2015 to 2020. They selected 3 types of information sources, including articles from journals, conferences, and dissertations. Their main goal was to introduce fog computing simulation tools and the challenge of developing such software applications.

In [17] the authors discussed the fog computing challenges and their open problems. They consider problems such as resource constraints, heterogeneous resources, dynamic and unpredictable nature of the

environment using resource management. Despite the resource management importance, there is no regular, comprehensive, and detailed review in the field of resource management methods in fog computing. They presented a systematic literature review (SLR) about approaches to resource management in fog computing based on a classical classification to identify advanced mechanisms. The classification has 6 primary areas, including programs, resource scheduling, performing tasks, resource allocation, resource supply, and load balancing. A Comparison of resource management approaches is done regarding essential factors, like performance criteria, techniques used, case studies, evaluation instruments, and strengths and drawbacks.

In [18], the authors presented an SLR-based scheme for examining the scheduling methods in fog computing. In their work, they presented scheduling algorithms and remarked on open issues. This study uses search methods to examine 100 papers published between 2012 and 2020. Then, they reached 36 final papers on scheduling approaches. In an SLR-based scheme, analysis of all available studies is not allowed. Thus, we ignored non-English, old, unreliable, and inaccessible works. The present review paper would be helpful for researchers for understanding the various dimensions of the reviewed subject properly. Nevertheless, it cannot consider all studies in this field because their number is very high and continuously increases. For future research, they want to propose a scheduling algorithm that can support dynamic environments and consider some evaluation criteria, like security and availability since most present algorithms only consider delay and cost as evaluation criteria.

In [19] the authors introduced a basic and general overview of fog computing architecture. They reviewed various resource and service allocation methods for addressing several important subjects, like delay, energy consumption, and bandwidth in fog computing. In comparison with other surveys, the present work presents an overview of advanced network programs and the main aspects of designing these networks. Moreover, the current research presents the Fog computing procedure and related challenges. This review paper discusses various architectures and determines the main research challenges. Fog computing can be used in many network applications because it can be considered the new version of predictive computations. Finally, they introduce some open research challenges and the essential design principles in their paper.

In [20] the authors considered the fog computing model as an option for IoT applications. This research examines concerns or challenges associated with fog computing for IoT. Researchers aim to address these problems associated with fog computing for IoT applications using a systematic literature review (SLR). They use the SLR scheme and apply search criteria to investigate an initial set including 439 papers from 2014 to 2017 and identify and review 17 studies related to this field.

These papers were organized into four main categories based on the challenges. This paper can help physicians and researchers understand concerns about fog computing and provide several useful views for future research directions. The scope of this paper is limited to the number of papers reviewed from the database. Based on the results of this study, they want to study precisely fog data dominance and business optimization techniques to use services of the IoT applications.

In [21] the authors presented a survey paper about the perspective, key features of fog computing, and new services and programs on the network edge. They stated that Fog must be a rich enough integrated platform to provide these new services. Also, it allows us to develop emerging services and new programs. In this paper, based on the fog computing characteristics, they introduced it as a suitable basis for supporting services and important IoT applications, smart networks, connected vehicles, smart cities, and generally wireless actuators and sensors. Researchers referred to three critical subjects in its research:

1) Fog architecture for massive computational infrastructure, storage, and communication devices.
2) Orchestration and management of fog nodes.
3) Services and programs supported by the Fog.

Fog Architecture and Its Characteristics

Fog computing can change how to provide services for customers for meeting IoT requirements. The fog infrastructure and services are extended both in the range of network and the cloud-to-Things continuum for allowing computational resources, which are located anywhere in this continuum, such as edge, cloud, or things for collecting these distributed resources and supporting programs [21] It is a high potential for task transfer from the cloud to the fog service providers, which are close to data sources or end-users. This can decrease the delay and bandwidth needed for data transmission to the cloud [22].

Fog computing is a new computational platform extending traditional cloud computing and services toward the network edge, providing communication, computation, storage, control, and service capabilities at the network edge.

The difference between the decentralized design and other common computational models is in terms of architecture. Fog computing means an integrated network concept that it stretches from the outer edges, which produce data, to where it is ultimately stored, whether in the cloud or the customer data center [23]

Accordingly, fog-computing framework development allows organizations to have more choices for data processing wherever it is more appropriate. In some applications, it is necessary to process data quickly. For example, when using a product connected device should immediately respond to the incident [7].

Fog computing can communicate with devices and analyze data at a low time.

This architecture can reduce the required bandwidth compared to when data should be sent to a data center or cloud. Therefore, it can also be applied in scenarios, in which there is no bandwidth for transferring data, so that information is processed somewhere close to its production location. As another benefit, users can add security features to a fog network. This can perform by dividing network traffic to the virtual firewall [24]. In the following, we present some of the strengths and drawbacks of fog computing.

A. Strengths of Fog Computing

- Reducing data transferred to the cloud
- Saving bandwidth
- Lowering the response time
- Increasing security through nearing data to the edge
- Supporting mobility

B. Drawbacks of Fog Computing

- Requiring hardware and more cost
- Requiring continuous access to the fog equipment

SLR Methodology

This section provides a summary of the SLR procedure to recognize, analyze and summarize the literature on a particular theme called SLR. The SLR attempts to find the original study to answer one or more questions [25], [26]. Through using a systematic literature review, the essential questions that can be asked about the research will be performed by considering the alternatives of the critical essential components. Then, by dividing the research into some main groups (domains) into the problem-solving approaches, the subsequent exploration string is defined [27], [27]:

(("service" OR "fog computing" OR "energy consumption" OR "energy saving" OR "energy-efficient" OR "energy management") AND (IoT OR Internet of Things)).

Fig. 2 shows the total papers selected from 5 valid databases, including IEEE, Wiley, Science Direct (Elsevier), Springer, and Taylor & Francis, based on the publisher and the number of papers published from 2013 to 2022. According to these figures, IEEE with 280 papers, about 18%, Springer with 258 papers, about 16%, Elsevier with 383 papers, about 24%, Wiley with 579 papers, about 36%, and Taylor & Francis with 96 papers, about 6%, were selected.

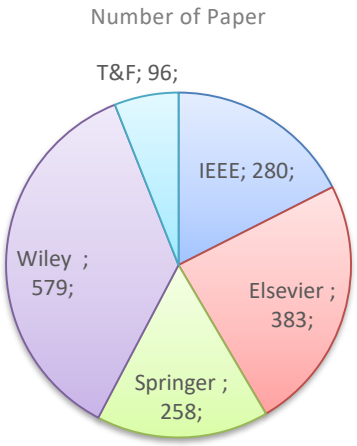


Fig. 2: A general comparison of the paper published by publishers.

A. Survey Goals and Research Questions

The paper presented in the SLR method, which is an evidence-based software engineering (EBSE), will answer the following Review Questions (RQ) to achieve the goals of this research. The objective of this chapter is to name the most important issues and difficulties in service, energy efficiency, energy-saving, and energy management in the fog computing field. This research is to direct the subsequent Research and Analytical Questions:

RQ1: Which scope and primary contexts are classified in service and Energy management in fog computing?

RQ2: What QOS parameters are utilized for evaluating the service and Energy management in fog computing?

RQ3: What assessment situations are utilized for the estimation of the service and Energy management in fog computing?

RQ4: What used tools for service and Energy management in fog computing?

RQ5: What is the significance of service and Energy management in fog computing?

RQ6: Which problems, future research directions, and challenges are identified concerning service and Energy management for future trends in fog computing?

These procedures can lead to comprehensive responses within the domain of this paper.

B. Search Query and Database Selection

The field of research is determined by selecting the most commonly used words to prepare our topic. Henceforth, seven keywords have been chosen, including "Service", "Fog", "Fog Computing", "Energy Consumption", "energy saving", "Energy Efficient" and "Energy Management". After various phases and using the results of our primary investigation as a pilot to analyze the coverage of the results, the inquiry is

characterized. To be specific, the query string is developed to include more keywords because the research in our model is not recovered by the essential inquiry, for instance, "IoT" OR "Internet of Things". To expand the domain of practical research, the search keywords and strings are just applied to the titles. The search is accomplished in October 2022, with a specified time range from 2013 until 2022.

C. Selection Criteria

The inclusion/exclusion scale for the final studies is applied after providing the analytical questions. Only the indexed ISI journal articles were investigated to limit the number of published papers. These articles are peer-reviewed papers on service management methods in the IoT field. For examining and addressing the referenced questions, 47 peer-reviewed papers are selected—these journal articles are presented in detail in next Sections. The selection ideology and assessment flowchart designed for the studies is shown in Fig. 3.

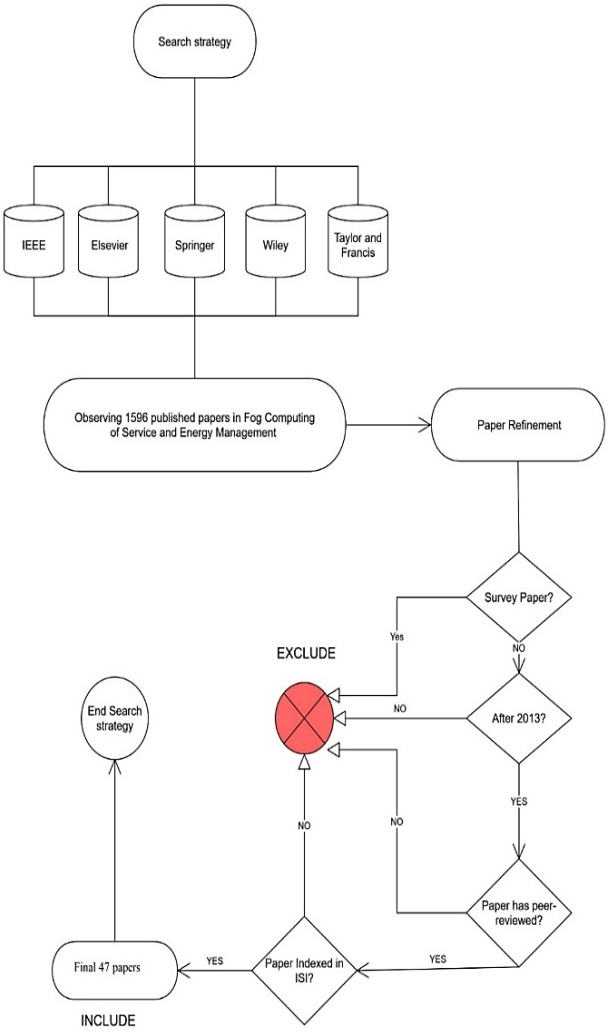


Fig. 3: Filtering approach of papers in this research.

There is some exclusion in our research, for example, short papers, low-quality studies, and non-peer-reviewed research (like predatory journals). The book chapters and white papers will be ignored because there is no research-based conversation and scientific data.

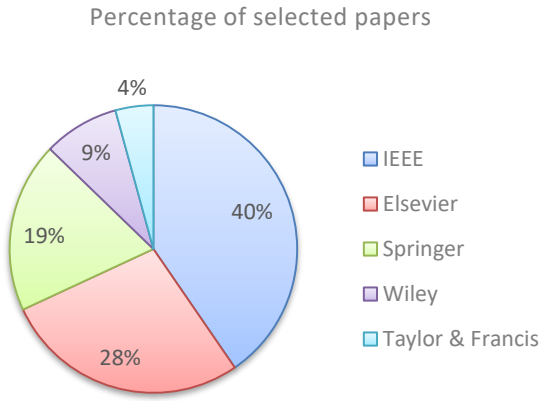


Fig. 4: The final percentage of selected papers based on the publisher in this research.

In this study, Fig. 4 displays the final papers selected from 5 valid databases, including IEEE, Wiley, Science Direct (Elsevier), Springer, and Taylor & Francis. Based on this figure, IEEE with 19 papers, about 40%, Springer with 9 papers, about 19 %, Elsevier with 13 papers approximately 28 %, Wiley with 4 papers, about 9%, and Taylor & Francis with 2 papers, about 4% were selected.

Service and Energy Management and Organization of the Study

IT management in today's environment is complex. cloud computing and fog computing introduces new path within IT and related organizations and industries [29]. The cloud computing can provide many services, in any case, how to deal with these services and ensure the quality of service has become one of the main Elements of its expansion. cloud computing services can be joined into complicated services and applications. cloud computing has been one of the challenges in improving cloud computing services for users and customers, and how to deal with these services and ensure the quality of (QoS) cloud services [30]. On the other hand, the worldview of fog computing has risen to be supplementary to cloud computing to simplify latency-critical, and applications with centralized bandwidth. In most cases, fog computing is deployed near the IoT devices/sensors and expands the cloud-based computing systems, storage accumulation, and facilities [31]. fog computing can be utilized as a viable stage for preparing power management as a service for various networking frameworks [31]. A fog service is a method for conveying quantity to clients by simplifying the results that clients need to accomplish. The services will deliver value to customers without the ownership of specific costs and risks [29].

One of the common goals of the research community, vendors, and suppliers is to design self-matchable solutions that can react to unpredictable workload oscillation and change the utility principles [32] The aspect of cloud computing services has created a noteworthy pattern of associations to make decisions about these services [33] Along these lines, the client can get crops with minimum price, high profitability, and business proficiency in a cloud environment [34] Because of the nature of fog Computing, which is distributed and has a diverse conditions, resource allocation is a significant issue. Many challenges need to be addressed to expand profitability and distribute appropriate asset tasks [35] In the real world, there is a limited resources and intensive and centralized computing applications to ensure a good experience [36]. The golden solution for this situation is task scheduling in distributed computing systems created by fog computing environments [37], [38]. The new advances in technology, cost, and scale of features have empowered us to manufacture computing devices with less power and performance than in the past [39] With different complicated digital physical vitality, the energy management frameworks should be actualized to be able to productively Monitor and deal with the task. To implement the energy management framework in the system, intelligence, interoperability, adaptability, and versatility are required [40].

Task scheduling will cause an energy consumption that is similarly performed between the poles of the system, and the organization of load network circulation happens consistently to extend the system lifetime [41] cloud suppliers need to address various key difficulties, such as finding some kind of harmony between the effectiveness of ideal vitality and the fulfillment of expanding requests and high-efficiency outlooks for clients [42].

The task scheduling in fog computing depends on whether there is an attachment between the scheduled tasks. It tends to be partitioned into two task scheduling: related and independent scheduling [43].

In this section, first, the presented 47 articles on Service and Energy Management are summarized. The next subdivision shows the various kinds of research in Service and Energy Management. Likewise, the various investigations will be thought about in various directions, for example, Main topic, Strong point of research, Deficiency of research, and new discovery. Fig. 5 shows the taxonomy of Service and Energy Management.

A. Architecture

In this section, we review two common categories in the Architecture section.

Design: In [44] the authors suggested an energy-efficient cross-layer-sensing clustering method (ECCM). Their algorithm applies the sensing-event-driven mechanism for putting fog nodes on the sensing layer and creating a robust virtual control node. In sensor networks, the cluster-based routing scheme is loaded on the fog layer, and the fog calculations use event nodes to achieve distributed clustering. Then, the optimized data aggregation routing has been made. Ultimately, the particle swarm optimization algorithm (PSO) is used to optimize the routing protocol by selecting an optimal set of nodes as cluster heads. The results of the proposed scheme show that it has the capability of optimizing the data aggregation process and improving network energy consumption and performance.

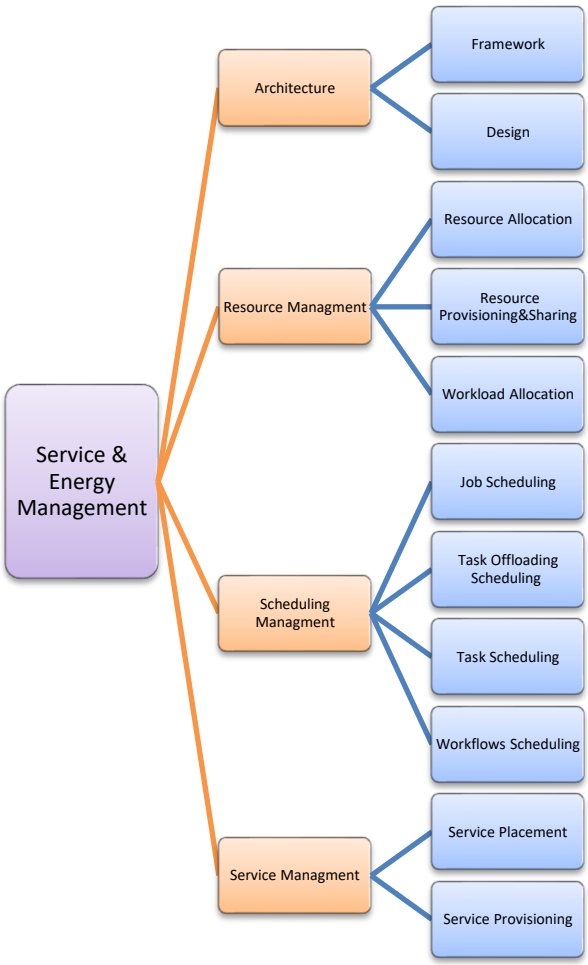


Fig. 5: The taxonomy of Service and Energy Management.

In [45] the energy efficiency of IoT resources was discussed as a critical subject in their research, and they proposed a low-consumption architecture for the IoT. It includes 3 layers: information processing, sensing and control, and presentation. In this architecture, sensors' sleep intervals are anticipated based on the residual energy of the battery and their previous use history. When the sensor nodes are in the sleep state, it is possible to use the predicted amount for strengthening the use of cloud resources by previsioning the allocated resources. This mechanism provides the optimal use of all IoT

resources. Experiments show that sensor nodes can save a significant amount of energy, and they can improve the use of cloud resources. The main characteristic of their model is the energy-information exchange between two layers. The proposed PA design is compared with EGF, SoT, OGL, and ECH. It shows that PA can reduce energy consumption compared to other schemes. Researchers in [46] conducted a study on an IPv6-based standard in fog computing networks aimed at improving scalability, latency, and performance, and a three-tier architecture. They designed a scheduling algorithm with 6TiSCH technology in fog computing, ensuring high reliability and low latency for emergency data transmission. They proposed a smart local discharge system to improve energy efficiency by rank-based Q-learning algorithm and to reduce latency by the fuzzy Bayesian learning method. The simulation results of their proposed design indicate its better results than the DIVA Priority and DeTAS methods in terms of energy consumption, latency, and other parameters.

Framework: In [47] the authors offered a 3-layer cloud-fog computing architecture to manage energy in adjustable micro-networks based on the dynamic thermal rating (DLR) constraints. It is possible to improve flexibility, reliability, and stability in power networks by decomposing large-scale networks into micro-networks. The problem is proposed as a mixed-integer linear optimization problem according to DLR constraints. The proposed scheme is compared with the backward-forward sweep scheme. The simulation results show that the proposed scheme has high performance to satisfy load requests and their constraints at any time, as well as the effective distribution of power in the network. In [48] the authors suggested a framework that calculates the tidal trust for the nodes, and based on prespecified values, malicious nodes are identified. The presented framework is assessed based on several IoT and FE devices. In the proposed framework, the rank and trust of IoT devices are dependent on the communication behavior of nodes. It is obtained via the security mechanism presented in this paper, which is more efficient than other methods. Identifying malicious nodes in early steps can improve the network performance in terms of efficient use of network congestion, network resources, packet loss rate (PLR), network power, and end-to-end delay.

B. Resource Management

In this section, we review three common categories regarding the studies conducted on resource management. The approaches are presented in Figure 5 as follows:

Resource allocation: This method is used to estimate resources or load resources to optimize them about the most appropriate resource.

Resource provisioning: This refers to the time of provisioning resources in the cloud and fog, which occurs by increasing and decreasing the duration as well as the time of providing/depleting real resources.

Resource sharing: Resource sharing is an approach to implementation and cooperation in the surface layers of the cloud or fog. These strategies are presented for sharing information, resources, and IoT programs.

Workload allocation: Transferring a large volume of IoT data to the cloud causes a long delay in workload processing, in which case workload allocation can be used for workload distribution according to energy and delay criteria.

Resource Allocation

In [49] the authors proposed a suitable scheduling algorithm with the 3-layer fog-to-cloud architecture for providing high quality and optimally using the fog-to-cloud sources. It has a successful performance in terms of service failure probability and delays. They enhanced the 3-layer fog-to-cloud architecture using the proposed design to decrease the delay in the data transmission process for delay-sensitive applications. Their simulation results confirm the performance and effectiveness of the scheduling algorithm proposed in this work. The proposed method outperforms existing algorithms such as the conventional cloud scheduling algorithm, the first-fit algorithm, and the random scheduling algorithm, but it does not consider energy consumption.

In [50] the authors addressed the efficient resource allocation problem in the fog environment, which is a fundamental problem in fog computing. They described how to allocate resources and how to put them in the virtual machine in the single-fog computing environment. For evaluating the presented algorithm in the fog environment, the architecture and the efficient resource allocation algorithm are executed in the CloudSim tool. According to the results, the presented method can allocate resources using an optimal method compared to the default resource allocation strategy. Their proposed scheme has a lower total processing time than other methods. The experiments show that the proposed framework improves the quality of a fog environment.

In [51] the authors discussed the joint computation and the communication resource allocation problem for the user's fog computing with the mixed task. They offered a mixed-task model for different types of computational tasks. This model supports binary loading and partial loading. Given the user satisfaction impact on the Fog computing services, their resource allocation goal was user-weighted energy efficiency (UWEE). Also, they utilized the concerned user mechanism (UCM) to draw the social characteristics of users. Then, to maximize UWEE, the resource allocation problem is designed as a mixed-integer nonlinear programming (MINLP) under the

constraint of users' satisfaction. However, it cannot be correctly solved due to binary depletion and traditional relaxation algorithms. a Lagrange-based resource allocation scheme called the augmented Lagrange method (ALM) is presented for solving this joint optimization problem iteratively, where AMSGRAD is used for accelerating convergence. The results of the simulation indicate the success of the ALM-based resource allocation scheme concerning UWEE.

In [52] the authors provided a PTPN-based resource allocation strategy for fog computing. It helps the user automatically select satisfactory resources from a group of pre-assigned sources. Their strategy considers comprehensively the time and cost spent to carry out a task and the trust evaluation of fog resources and users. Based on the fog features, they create the PTPN task models in fog computing. In the proposed scheme, they suggested an algorithm predicting the task completion time. Also, they consider an evaluation scheme to calculate the validity of the fog source. The dynamic fog source allocation algorithm has a higher throughput compared to static allocation strategies in terms of concerning cost and time. However, researchers ignore energy consumption in this study.

In [53] the authors addressed the challenges of medical cyber-physical systems (MCPS) and long-delayed and unstable links between the cloud data center and medical devices. For dealing with this issue, they presented “Mobile edge cloud computing” or “Fog computing” as a solution. The integrated fog computing and MCPS introduced the FC-MCPS scheme, firstly formulating the problem as a mixed-integer non-linear linear program (MINLP) and then formulating it as a mixed-integer linear programming (MILP). To solve computational complexity, they suggested a two-phase linear programming-based heuristic algorithm (LP). In the presented scheme, the results generate an accurate and proper solution. As a result, it outperforms a greedy scheme. In addition, the proposed method is cost-efficient.

Researchers in [54] studied the connection of IoT devices generated requests and concerns about energy consumption and delays in fog networks. To solve this problem, they formulated the problem and used the mesh adaptive direct search algorithm (MADS) algorithm to find the optimal results. The proposed algorithm was more efficient than OOA and ESA algorithms. The simulation results show that the performance of MADS can be similar to the two algorithms but their proposed method has less computational complexity.

Resource Provisioning & Resource Sharing

In [55] the focus of the authors is on the resource allocation problem between fog nodes and mobile users to consider cost efficiency. Their purpose is to allocate the load for users and reduce costs to meet computational and communication constraints. They considered three conditions to allocate resources to several users in fog computing. These conditions are the single-version resource allocation, resource allocation with several duplications of the transmission cost model, and resource allocation with several duplications of various transmission costs. Two models have been discussed in the proposed method: the different transmission cost model and the transmission cost model. Users can load multiple versions with fixed transmission costs for the transmission cost model. So, a suitable greedy solution is proposed. The transmission cost is associated with the distance between pairs of fog nodes for the different transmission models. Therefore, they suggested a non-adaptive algorithm. The performance of the proposed algorithms is evaluated based on two real databases. The results confirm the efficiency and effectiveness of the proposed algorithm.

In [56] the authors presented a computational fog structure and a crowd-funding algorithm to integrate additional resources into the network. In the proposed algorithm, they proposed an encouraging mechanism to incentivize owners with more resources to share their resources with the resource set and supervise and support resources during the active tasks. The results and the simulation indicate the effectiveness of the presented encouraging mechanism in reducing the violation of SLA and accelerating the completion of tasks.

Researchers in [57] discussed the IoT layer, fog computing, and the choice of close source or sources. They examined resource discovery. They used hidden Markov chain learning in their proposed method and compared it with TOPSIS, VIKOR, and SAW methods. Compared with the existing methods, the efficiency of their proposed method led to a reduction in energy consumption.

workload Allocation

In [58] the authors addressed finding the node’s location in the Fog. In this regard, their purpose is to solve the problem of the fog node’s location for users who use mobile phones with limited batteries. Therefore, they provided a solution that supports limited energy sources and computations with low delay. They used the mixed-integer linear programming (MILP) formula in the proposed solution to solve the problem. Also, they provided innovative solutions to solve large-scale problems. The results obtained from a real mobility database indicate that the innovative solution is more accurate than the MILP formula. As a result, it can significantly save energy for end-users.

In [59] the authors investigated the delay and energy-efficient load allocation problem in the IoT-Edge-Cloud computing system. They adjusted a load allocation

problem based on delay. Their scheme is the optimal load allocation among local edges, neighboring servers, and clouds to minimize energy consumption and reduce delay. In their proposed method, they applied the delay-based workload allocation (DBWA) algorithm, the drift theory, and the Lyapunov penalty to solve this problem. The simulation results show that the proposed method improves energy efficiency and reduces delay in an IoT-Edge-Cloud system.

Researchers in [60] provided a collaborative scheme on the challenge of service delivery and the long distance to the computing resources. In this solution, a three-tier architecture is used to overcome the mentioned limitation. They proposed a trust-based offloading method, the efficiency of which is evaluated in comparison with SSLBA, NFA, and RWA methods. The simulation results show that the proposed method can significantly reduce latency and improve the service response rate.

C. Scheduling Management

In this section, we review four common categories regarding the studies conducted on scheduling management. The approaches are presented in Fig. 5 as follows:

Workflow scheduling: Transferring the workflow to the cloud/fog computing environment and using different services to facilitate the implementation of the workflow.

Job scheduling: This refers to allocating users' tasks to virtual machines for execution. From the users' perspective, an appropriate scheduling algorithm should be able to perform the required tasks within the shortest time.

Task scheduling: This refers to allocating tasks to processing resources in such a way that some system performance parameters such as execution time can be optimized.

Task offloading scheduling: Mobile devices are faced with limitations such as limited battery life, low processing power, and limited storage space, which are effective in improving the quality of services provided to customers. To overcome these limitations, tasks that require processing and storage should be transferred to the cloud or other mobile devices around the user or to a combination of both.

Job Scheduling

In [61] the authors discussed fog computing and better responsiveness to all users' demands and computational requirements. They develop and simulate the fog computing system based on networks, smartphones, containers, and clouds in their paper. Their purpose is to check the possibility of using any available resources to reduce the total costs. It is performed with a Bag-OF-Tasks load model. The simulation results show that it reduces costs without increasing the average response time. In this study, energy consumption is not considered.

In [62] the authors suggested a new optimization approach called BLA8. It aims to solve the task scheduling problem in the fog computing environment. This method seeks to distribute some tasks among all fog nodes optimally. Its main goal is to make an appropriate tradeoff between the CPU runtime and the assigned memory to improve mobile users' needs for Fog computing services. The performance evaluation results indicate that their proposed scheme outperforms traditional algorithms, like PSO and GA, in terms of CPU runtime and the allocated memory. However, they ignore energy consumption in this study.

In [63], the authors presented a model to schedule the requests of the IoT service to minimize the overall service request delay. In this model, they used integer programming for solving the optimization problem. Furthermore, they introduced a customized genetic algorithm as an innovative method for scheduling requests of IoT and minimized the delay. They tested the customized genetic algorithm in a simulation environment, with consideration of the environment's dynamic nature. They evaluate the customized genetic algorithm and compare it with three schemes, including priority-strict queuing (PSQ), waited-fair queuing (WFQ), and round-robin (RR). It outperforms others in terms of delay and meeting the requests deadlines.

In [64] researchers tried to solve the problem of energy demand optimization. They proposed a scheduling method based on fog and cloud computing, which focused on reducing electricity consumption demand during peak hours. In the proposed method, the fog nodes provided consumers' priorities, which makes resource allocation more efficient by creating interaction between nodes. The use of node allocation reduces scheduling delays.

Task Offloading

In [65] the authors examined the loading problem in dynamic computing on the IoT-Fog systems. Their research assumes that channel status information can be fully obtained by the depletion agent. They presented a partially visible depletion scheme, enabling the IoT device to make an optimal depletion decision using incomplete channel status information. To minimize energy consumption and delay, the optimization issue has been formulated using the partially observable Markov decision process (POMDP). For finding the optimal loading solution, an offline algorithm based on the deep recurrent Q-network (DRQN) has been created. In the proposed POMDP solution, a DRQN-based offline algorithm is created. It combines LSTM and DQN. Compared to other depletion schemes, as shown by

numerical results, their presented scheme can effectively decrease the energy consumption in the IoT devices and reduce delay during processing the computational tasks.

In [66] the authors suggested a task-loading design based on an enhanced contract NET protocol and the beetle antennae search algorithm in fog computing networks. In this problem, the distribution of the task nodes and fog nodes is done uniformly in a circular area with R radius. The responsibility of the task node is to divide the task into sub-tasks and send them to the fog nodes. The proposed scheme has been applied for reducing the task node cost. The proposed algorithm combines the beetle antennae search algorithm and GA. This method only focuses on reducing costs and does not address other parameters, like energy consumption.

In [67] the authors offered a four-layer architecture for determining the decision maker for task depletion. In this study, the issue is formulated as a population (evolutionary) game, which is solved with Replicator Maynard dynamics. In this study, optimization objectives are time and energy consumption. They emphasize using realistic parameters and values to simulate the proposed scheme. The findings show the practicality of their design in reducing main traffic.

In [68] authors considered a three-layer fog computing architecture. Mobility in user equipment is determined according to sojourn time in each Fog computing node. It can be expressed as an exponential distribution. The purpose of the researchers is to maximize the efficiency of user equipment and optimize decisions of loading and computational resource allocation to reduce the migration probability. They formulated this problem as an MINLP and divided the problem into two parts: (1) Loading tasks and (2) Resource allocation. They introduced a fog computing node selection scheme based on the Ginni coefficient, which is called GCFSA, to obtain an optimal off-loading strategy, as well as a distributed resource optimization algorithm based on the GA called ROAGA for solving the resource allocation problem. Their simulation results show that their scheme outperforms other basic algorithms and can obtain a quasi-optimal performance. In this simulation, they focus on reducing migration.

In [69] the authors presented a distributed learning scheme to minimize the fog computations' average cost if there is no knowledge about random traffic in non-DTN application scenarios. They presented a fully distributed learning approach to minimize the average cost and time of fog computations. Stochastic gradient descent is used to separate optimal operations between time slots to create a distributed evolutionary heuristic for separating and achieving semi-optimal approximation. Online learning can reduce the drop-in optimality caused by distributed scheduling. Their simulation results show that the proposed distributed learning is better than other schemes in terms of operating power and energy efficiency.

Researchers in [70] examined a multi-objective problem to optimize task completion time and energy consumption by combining GA and PSO algorithms. They proposed a task offloading plan to decide on offloading, select appropriate fog nodes, and allocate computing resources. The proposed method had better performance than GA, PSO, local computing, random offloading, and uniform offloading algorithms in terms of energy consumption and overall offload overhead.

Task Scheduling

In [71] the authors designed a model to solve the multi-objective task scheduling problem in fog computing. Their scheme is an adaptive multi-objective optimization method to schedule tasks. In the adaptive multi-objective optimization task scheduling scheme (AMOSM), they have considered the total runtime and the cost of task resources in the fog network as the resource allocation optimization objectives. Their experiments show that the presented scheme had a better performance compared to other methods concerning total runtime, cost of resources, and loading.

In [72] the authors addressed the complex task scheduling problem. In addition, they consider the consumed energy to reduce energy consumption if a mixed deadline condition is met in the IoT applications. They adjusted a limited optimization in the cloud-fog environment for solving the task scheduling problem. This problem can be solved using the laxity and ant colony system algorithm (LBP-ACS). In their proposed scheme, a hybrid task scheduling strategy is considered. It includes the priority of a task and its deadline. To manage the delay sensitivity in a task, the Laxity-based priority algorithm seeks to build a task scheduling sequence with proper priority. Furthermore, the limited optimization algorithm applies the ant colony algorithm to minimize total energy consumption. They compared their proposed method with other algorithms, the results of which indicate the effectiveness of the presented algorithm in reducing energy consumption to process all tasks. It also can ensure the appropriate scheduling length and decrease the failure rate of the scheduling of tasks with different deadlines.

Researchers in [73] conducted a study on Elastic Optical Networks (EONs) in the underlying basic tiers. Their main focus was on solving the traffic problem of fog services and reducing excess energy consumption. They proposed an Energy-efficient Deep Reinforced Traffic Grooming (EDTG) algorithm. They extracted features that they implemented with the Advantage Actor-Critic (A2C) algorithm and an artificial neural network (ANN). The results show that the proposed algorithm can significantly

reduce energy consumption compared with the two DRL and SGA algorithms.

Researchers in [74] proposed an efficient method of task scheduling in a heterogeneous virtual cloud by focusing on the energy consumption reduction problem. Their proposed method uses a logical balance method between task scheduling and energy saving. The mechanism of their proposed method is such that they will first have an initial schedule to reduce the execution time and then re-plan to find the best execution place in due time with less energy consumption. The proposed EPETS method has significantly better performance than energy-efficient scheduling methods such as RC-GA, AMTS, and E-PAGA in terms of energy consumption.

Researchers in [75] developed an alternative technique for IoT requests called AEOSSA in a cloudy environment. The AEOSSA method uses a combination of AEO and SSA algorithms to solve the task scheduling problem. The performance of the AEOSSA approach designed to solve the scheduling problem was compared with five traditional metaheuristic techniques. The simulation results showed that the proposed method had better throughput and makespan than the other 5 methods.

Researchers in [76] proposed an IEGA algorithm to solve the problem of scheduling tasks in fog computations. The mechanism of the proposed method consists of two steps: first, the mutation rate and the crossover rate are set to find the optimal combinations, and second, several solutions are mutated based on a certain probability to discover a better solution and not get stuck at local minima. The proposed method was compared with five evolutionary optimization algorithms. The proposed method was shown to be superior to other algorithms in terms of energy consumption, makespan, and several other parameters.

Workflows Scheduling

In [77] the authors presented a three-layer architecture, including Fog, cloud, and consumer layers. They presented a meta-heuristic algorithm called the improved particle swarm optimization with levy walk (IPSOLW) for load balancing. This algorithm combines PSO and LW. Users send their requests to the fog servers and then receive services. When the fog layer is damaged, the cloud is used for storing the consumers' records and providing services to them. Finally, a comparison is made between their algorithm and available algorithms like BAT, PSO, BPSO, CLW, and GA. They evaluated some parameters, including response time, cost, and processing time. Experiment results show that their proposed algorithm outperforms other algorithms.

In [78] the researchers presented a hybrid architecture for the dynamic scheduling of several tasks in the real-time IoT. In traditional approaches, the IoT task processing is conducted on the fog layer; whereas, in their approach, it is attempted to schedule computational tasks with low communication necessities in the cloud and tasks with compact communications and low computational requirements in the Fog. Their scheme considers the communication cost during the scheduling process, which is due to data transferred from devices and sensors in the IoT layer to the fog layer. The performance of their scheme is evaluated by simulating an unaware cloud strategy. The simulation results show that their proposed scheme has a lower deadline compared to the base policy.

A. service management

In this section, we review two common categories regarding the studies conducted on service.

Service Placement

In [79], the authors provided a method, which is based on the monitoring, analysis, decision-making, and execution (MADE) methodology, to order the IoT services. This scheme operates autonomously. In this scheme, the first existing resources and the status of program services are controlled at runtime. In the next step, the requested services are prioritized according to the deadline of program services. Then, the evolutionary strength Pareto II algorithm is used to decide on ordering program services as a multi-objective optimization problem. Finally, the decisions taken in previous phases are implemented in the fog environment. Their proposed scheme outperforms MOPSO and NSGAII algorithms in terms of various criteria such as service latency, fog utilization, and cost.

In [80] the authors described the various computing paradigms, like cloud computing, fog computing, and their combination, which is known as F2C computing. Their purpose was to study the benefits of fog computing and F2C computing and the delay support pattern in the Fog. They believed that F2C systems have a suitable performance because they can provide and implement distinct strategies for evaluating distributed services in F2C. The results show that the distributed implementation of sources in F2C has many benefits concerning service response time and main network load. In this study, the results of the cloud-based resource allocation scheme (CL) are compared with the three placement strategies, including FF, BF, and BQ, for four modes, including network delay, processing delay, general delay, and main network load. However, they do not address heterogeneous sources of the Fog.

In [81] authors examined the services on mobile networks equipped with fog computing. They suggested a QoS-aware scheme based on existing delivery methods to support vehicle services in real time. This paper presented three designs, including without any migration services,

with migration services of HANDOVER, and the proposed design. The two first designs have their weaknesses. Thus, the third design has been introduced. The main idea of the third design is to combine two strategies to minimize migration overhead and maintain end-to-end performance at an acceptable level to meet QoS requirements. In their research, the authors have conducted a case study based on the real vehicle mobility pattern in a small European country. The proposed design is evaluated based on three criteria, including delay, reliability, as well as migration costs.

In [82] the authors investigated the SFC migration problem/reconstruction developed by user motion in cloud-fog computing environments. In the first step, they formulated the SFC migration problem as an integer linear programming. Then, they suggested two SFC migration approaches, including the minimum number of VNF migration strategies and a two-phase migration strategy. The two-phase proposed migration strategy is simulated in the cloud-fog computing environment. It can improve the configuration cost, reconstruction success rate, migration time, and failure of the proposed algorithm compared to basic algorithms.

In [83] the authors examined the resource supply problem during the use of fog computing resources. They proposed a conceptual, computational framework. Then, they formulated the service replacement problem for IoT applications in fog resources as an optimization problem. The authors consider heterogeneous resources and applications in their paper. They proposed a genetic algorithm as evolutionary solution to solve the optimization problem. The simulation results show that the implementation of this service can reduce communication delays in the fog network.

Researchers in [84] discussed the challenges of the geographic distribution of nodes for the proper management and processing of requests in fog computing. They provided a two-tier fog framework. They formulated the problem with the EGA algorithm considering the parameters to reduce the service time, costs, and energy consumption and thus ensure the QoS of the IoT system. The results of the proposed method had better performance than DEBTS, DMS, FIRST-FIT, branch and bound, and GAPSO algorithms in terms of service cost, energy consumption and service time.

Researchers in [85] provided a trust management system (TMS) regarding the security of loading and offloading requests to the cloud computing layer. In the proposed system, the service requester first checks the trust condition of the service provider and then sends the requests. In addition to QoS issues, it also provides QoS. Fuzzy logic is used in their proposed method to ensure the security of services. The proposed method has better performance than a method without TMS in terms of Delay and Throughput.

In [86] researchers proposed a solution using MAPE-K methodology in a fog environment and the Whale Optimization Algorithm (WOA) to solve the service placement problem. In this solution, operational capacity and energy consumption were considered to be the main objective of their research. This method is implemented on a three-layer architecture to show the interaction between the IoT device and fog layers. The results of their proposed solution reduce resource consumption, service delay, and energy consumption compared to other meta-heuristic methods.

In [87] researchers used a genetic algorithm to solve the service placement problem, reduce the delay of programs in the cloud-fog environment, and use the network. In the proposed method, they defined a penalty parameter to reduce the delay. In the results of their proposed method, delay, network use, energy consumption, and cost were improved to an acceptable extent.

In [88], researchers proposed a computational framework to solve the service placement problem in the cloud-fog environment and optimize the IoT services. They formulated the service placement problem as an automatic planning model considering the heterogeneity of programs and resources. They proposed the use of the PSO algorithm to solve the problem of IoT service placement to maximizing the use of fog resources. Their simulation was based on various QoS criteria, which led the PSO-based method to exhibit better performance than other advanced methods.

Service Provisioning

In [89] the authors first suggested FOGPLAN; and then they introduced the QDFSP framework. The purpose of this framework is to dynamically provide QoS-aware fog services. QDFSP is related to the dynamic deployment of application services in fog nodes or the release of application services, already deployed in fog nodes for meeting the lowest cost, delay, and QoS prerequisites of applications. FOGPLAN is a practical framework, which can operate with no hypotheses and with the lowest information concerning the IoT nodes. The authors used an integer nonlinear programming formula as an optimization issue. Also, they presented two greedy algorithms, namely Min-Cost, and efficient Min-Cost, to address QDFSP periodically. In larger settings, Min-Cost generally has less runtime, meaning that it is faster. The Speed of Min-Viol is lower than Min-Cost, but Min-Viol has fewer delay violations and an average service delay compared to Min-Cost. Finally, QDFSP cannot be used for solving the optimization problem in a periodical manner, especially for large networks. This research mainly

focuses on cost and delay and does not examine large-scale heterogeneous networks.

In [90] authors investigated the multimedia fog computing support algorithm, synchronization aspects during the use of resources in large-scale systems, and their ability to guarantee the QoS requirements. They added fusion technology of privacy protection and service location to large-scale multimedia applications. contrary to computing, energy supply, network access capacity, storage, and other factors of fog nodes, the fusion technology has the ability to obtain location data of the positioning service in real-time mobile terminals. The proposed method evaluates the time cutting, hardware, bandwidth, and other network resource balancing technologies, and the Fog computing support algorithm. It can improve the security and performance of the multimedia fog computing system. They simulated the FCS-FLSPP algorithm. Then, its results are compared with the MBCS-VSD algorithm in terms of different parameters, including the high use of multimedia system resources, and their ability for load balancing in real-time. It has a better performance than the MBCS-VSD algorithm.

Discussion

This section provides some systematic reports on the planned explanatory inquiries, which are as follows:

RQ1: Which scope and primary contexts are classified in service and Energy management in fog computing?

Table 2 present a comparison of the Service and Energy Management based on the presented taxonomy in the previous Section. There are four scopes are considered, including Architecture, Resource Management, Scheduling management, and finally Service Management. Scheduling Management has the greatest portion of the area with 38% usage in the literature. Furthermore, Resource management has about 26% usage in fog computing. Fig. 6 represents the amount of service and energy management domains.

Table 2: Categorization of the service and energy management approaches in the fog computing field

Category	Research	Strong point of research	Deficiency of research	New discovery
Architecture	[44]	Low Energy consumption	Not multiple objective	**Algorithm**
	[45]	Low energy Consumption Low response time Low resource utilization	data generated and considered in this study by each sensor was few. just a 2-hour experiment was considered	**Architecture**
	[46]	Low energy consumption Low delay Low response time Low transmission efficiency High Throughput	Not considering assessment of computation complexity	**Architecture** **Algorithm**
	[47]	Low latency High data Transmission Speed Time	very mathematical formulation Not compared to conventional networks	**Framework**
	[48]	Low Response Time Low resource utilization Low Cost Low Execution time	Only Simulation with 3 VM Machine	**Framework**
Resource Management	[49]	Low response time	Not consider energy consumption	**Architecture**
	[50]	Enhances the Efficiency Resource Management	Only 1 parameter were simulated	**Algorithm**
	[51]	less computation complexity	very mathematical formulation	**Algorithm**
	[52]	Low Completion Time Low price cost heterogeneous network	Not consider energy consumption very mathematical formulation	**Algorithm** **Strategy**
	[53]	Low cost	Only cost parameter were simulated	**Algorithm**
	[54]	Low Energy consumption low computational power very little complexity	Experiment with fewer IoT nodes and FOG devices	**Algorithm**
	[55]	low Replication cost low Transmission cost	Only 1 parameter (cost) were simulated very mathematical formulation	**Scheme**
	[56]	Low completion time	Not consider resource utilization Not consider energy consumption	**Algorithm**
	[57]	Low Energy consumption Resource Efficiency	Not consider delay	**Algorithm**
	[58]	Low Energy consumption	Very mathematical formulation not consider heterogeneous end-user devices	**Algorithm**
	[59]	Low Energy consumption Low delay	Not compared to popular algorithms	**Algorithm**
	[60]	Reduced Service Latency	support only static IoT devices	**Algorithm**

Category	Research	Strong point of research	Deficiency of research	New discovery
		Service Response Rate		
Scheduling Management	[61]	Low response time Low Utilization of Resources Low cost	Not consider energy consumption	**Model**
	[62]	Low execution time High Allocated memory Low Cost	Not consider energy consumption	**Algorithm**
	[63]	Low Latency Low Runtime Improved meeting of Requests Deadlines	Not multiple objective	**Algorithm**
	[64]	Low delay Low energy consumption Low cost	Not considering other qos metric	**Algorithm**
	[65]	Low delay Low Energy Consumption	Simulator engine was not mentioned	**Algorithm**
	[66]	Low Average Cost	Only cost parameter were simulated	**Algorithm**
	[67]	Low delay Low Energy Consumption Improved Convergence Time	Only offloading parameter were simulated	**Architecture**
	[68]	Low Migration Times Improved the Revenue of UEs Low Energy Comparison	Not considering other qos metric	**Algorithm**
	[69]	Low Energy Comparison Low Delays Low Running Times Low Time Cost	very mathematical formulation	**Algorithm**
	[70]	Task Completion time and Energy Consumption.	Not consider other qos parametrs	**Algorithm**
	[71]	Low Delay Low Execution Time Low cost	very mathematical formulation not implemented in physical environment	**Algorithm**
	[72]	Low Energy Consumption Low Scheduling Length of the Task	Not consider delay	**Algorithm**
	[73]	Reduce Energy consumption	simulated only with 14 Nodes and 21-link Not consider other Qos parameters	**Algorithm**
	[74]	Low Energy Consumption Improve Performance	execution time	**Algorithm**
	[75]	Makespan time Throughput	Not consider other Qos parameters	**Algorithm**
	[76]	Makespan, Flow Time Fitness Function Carbon Dioxide Emission Rate Energy Consumption	Not considering scaling large.	**Algorithm**
	[77]	Low response time Low Average processing time. Low Cost.	Not consider energy consumption	**Architecture.** **Algorithm**
	[78]	High Tasks Executed Low Cost Low Deadline Miss Ratio	Not consider communication cost by the transfer of data in fog layer	**Algorithm**
Service management	[79]	Low Utilization of resources Low Latency Low Cost	Not considering energy consumption Not considering their priority policies	**Algorithm** **Framework**
	[80]	Low Delay Low Response Time	heterogeneous fog resources Only delay parameter was simulated	**Algorithm**
	[81]	Low Latency Low Migration Cost High Reliability	Not considering other qos metric	**Scheme**
	[82]	Low cost Low Running Time Low Migration Time Low Down Time	Not considering scaling large.	**Algorithm**
	[83]	Low Response Times Low Service Execution Delays Low Resources Utilization	heterogeneous	**Method**

Category	Research	Strong point of research	Deficiency of research	New discovery
	[84]	Low Service cost Low Energy Consumption Low Service Time Low Resources Utilization	Docker settings are not described	**Algorithm**
	[85]	High Throughput Low Delay	Low range low number of clients	**Method**
	[86]	Low Resources Utilization Low Energy Consumption Low Delay	Only Simulation with 5 and 10 service type in fog nodes	**Algorithm**
	[87]	Low Resources Utilization Low Energy Consumption Low Latency Low Execution Time Cost	Only Simulation with maximum 4 fog node and 8 End Devices	**Algorithm**
	[88]	Low Execution Time Average Waiting Time Low Failed Services	Only Simulation with 5 service type and 5 fog nodes. Not considering energy consumption.	**Algorithm** **Framework**
	[89]	Low cost Low Delay	Not considering scaling large.	**Framework** **Algorithm**
	[90]	Low Resources Utilization Low average waiting time of Service	real-time performance is slightly	**Algorithm**

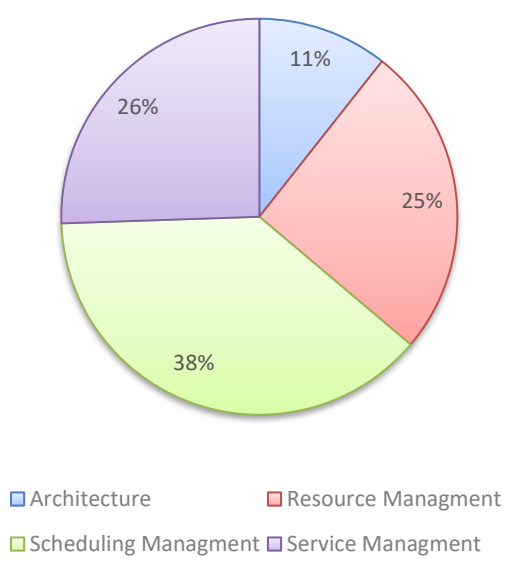


Fig. 6: Percentage of Service and Energy Management domains.

RQ2: What QOS parameters are utilized for evaluating the service and Energy management in fog computing?

Table 3 lists the QoS parameters used in fog computing.

This table introduces six parameters: response time, latency/delay, energy consumption, cost, resource utilization, throughput, and execution time. As shown in Fig. 7, the cost parameter has the highest percentage (i.e., 51%).

After the cost parameter, the two parameters, namely energy consumption & execution time, have a higher percentage than others (i.e., 42%). In this field, the percentage of resource utilization and latency/delay is equal to 33% and 30%.

Response time has the least percentage (i.e., 23%). Hence, it seems that it has been less investigated in the research literature.

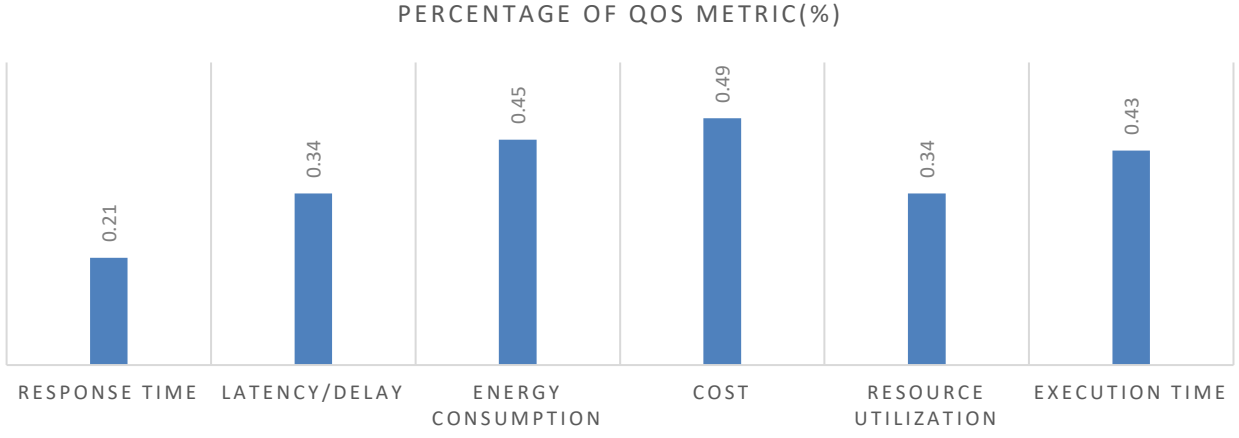


Fig. 7: The QoS parameters evaluated in this research.

Table 3: The QoS parameters reviewed in this research

Reference	Year	Response Time	Latency/delay	Energy consumption	Cost	resource utilization	Execution time
[44]	2019			✔			
[45]	2015	✔		✔		✔	
[46]	2021	✔	✔	✔			
[47]	2020	✔		✔	✔	✔	
[48]	2020	✔			✔	✔	✔
[49]	2020	✔					
[50]	2020				✔		
[51]	2020			✔			
[52]	2017				✔		✔
[53]	2015				✔		
[54]	2021			✔			
[55]	2020				✔		
[56]	2017						✔
[57]	2021	✔		✔		✔	✔
[58]	2020			✔		✔	
[59]	2019		✔	✔			
[60]	2021		✔				✔
[61]	2020	✔			✔	✔	
[62]	2018				✔	✔	✔
[63]	2020		✔				✔
[64]	2022		✔	✔	✔		
[65]	2020		✔	✔			
[66]	2020				✔	✔	
[67]	2020					✔	
[68]	2019			✔	✔		
[69]	2018		✔	✔	✔		✔
[70]	2021			✔			✔
[71]	2020		✔		✔		✔
[72]	2019			✔			
[73]	2021			✔			
[74]	2021			✔		✔	✔
[75]	2021						✔
[76]	2021			✔			✔
[77]	2019	✔			✔		✔
[78]	2019				✔		✔
[79]	2020		✔		✔	✔	
[80]	2018	✔	✔				
[81]	2019		✔		✔		
[82]	2019				✔		✔
[83]	2017	✔	✔		✔	✔	✔

Reference	Year	Response Time	Latency/delay	Energy consumption	Cost	resource utilization	Execution time
[84]	2021			✔	✔	✔	✔
[85]	2021		✔		✔		
[86]	2022		✔	✔		✔	
[87]	2022		✔	✔		✔	✔
[88]	2022						✔
[89]	2019		✔		✔		
[90]	2019				✔	✔	
Count		10	16	21	23	16	20

RQ3: What assessment situations are utilized for the estimation of the service and Energy management in fog computing?

Table 4 investigates different parameters, such as simulator, simulation environments, algorithm, and architecture used in papers. In service and energy management, 37 papers have only simulated their scenario; 7 papers have implemented their ideas for service management in the real environment. In addition, 3 papers have implemented a part of their work; they have done simulations to analyze and compare its results with others.

Fig. 8 represents the simulation of papers on energy consumption and service management approaches. Fig. 8 indicates the statistical percentage of the schemes. It is observed that 81% of the papers have been simulated using different tools. Furthermore, 15% of the papers have been implemented in the real environment, and 4% of them are simulated with other data; they belong to the third category.

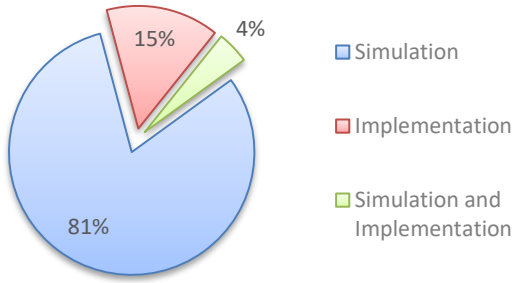


Fig. 8: Simulation and implementation used in the research literature.

The energy consumption and service management approaches consider an important subject, including architecture and framework, along with algorithms. This is depicted in Fig. 9. This figure displays the statistical percentage of the schemes. It can be seen that researchers present a framework or architecture in 30% of the papers. Also, 70% of the papers have no architecture or framework.

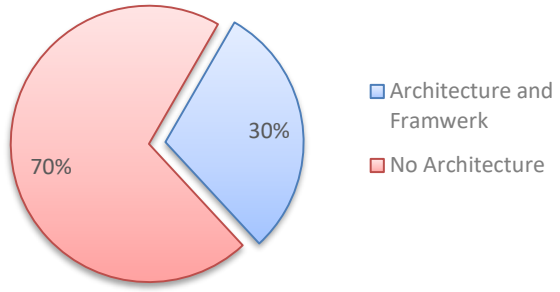


Fig. 9: Percentage of representation of architecture in the research literature.

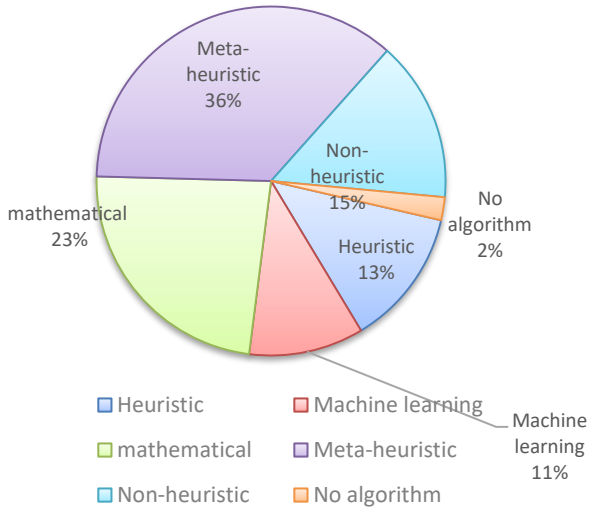


Fig. 10: Different percentages of various algorithms in the research literature.

Fig. 10 depicts algorithms and tools used in energy consumption and service management approaches. The statistical percentage of the methods is shown in Fig. 10. It can be seen that 36% of research literature includes papers, which use Meta heuristic methods. After that, mathematical schemes are included in 23% of the research. Finally, 2% of the papers do not use any algorithm. They have often presented a framework or architecture.

Table 4: Investigating the parameters of simulators, algorithms, and architecture in this research

Number	Data base	Year	Reference	Simulator	Evaluation	Algorithm	Algorithm type	Architecture
1	IEEE	2019	[44]	MATLAB	S	✔	Meta Heuristic	
2	IEEE	2015	[45]	HADOOP	I		Mathematical	✔
3	IEEE	2021	[46]	NS3	S	✔	Machine learning	✔
4	IEEE	2020	[47]	NM	I		Mathematical	✔
5	SPRINGER	2020	[48]	NS2	S		Non-Heuristic	✔
6	SPRINGER	2020	[49]	CLOUDSIM	S	✔	Non-Heuristic	✔
7	WILEY	2020	[50]	CLOUDSIM	S, I	✔	Non-Heuristic	✔
8	IEEE	2020	[51]	NM	S		Non-Heuristic	✔
9	IEEE	2017	[52]	NM	S	✔	Mathematical	
10	IEEE	2015	[53]	NM	S	✔	Heuristic	
11	WILEY	2021	[54]	NM	S	✔	Mathematical	
12	ELSEVIER	2020	[55]	NM	S	✔	Heuristic	
13	ELSEVIER	2017	[56]	HADOOP	I	✔	Machine Learning	
14	SPRINGER	2021	[57]	CLOUDSIM	S	✔	Mathematical	
15	IEEE	2020	[58]	PYTHON	I	✔	Heuristic	
16	IEEE	2019	[59]	MATLAB	S	✔	Mathematical	
17	ELSEVIER	2021	[60]	NM	S	✔	Mathematical	
18	ELSEVIER	2020	[61]	JPFF	I	✔	Non-Heuristic	
19	T & F	2018	[62]	C++	S	✔	Meta Heuristic	
20	ELSEVIER	2020	[63]	NM	S	✔	Meta Heuristic	
21	IEEE	2022	[64]	NM	S	✔	Mathematical	
22	IEEE	2020	[65]	NM	S		Machine learning	✔
23	SPRINGER	2020	[66]	NM	S	✔	Meta Heuristic	
24	SPRINGER	2020	[67]	PHYTON, MATLAB	S, I		Meta Heuristic	✔
25	IEEE	2019	[68]	NM	S	✔	Meta Heuristic	
26	IEEE	2018	[69]	NM	S	✔	Machine Learning	✔
27	ELSEVIER	2021	[70]	NM	S	✔	Meta Heuristic	
28	IEEE	2020	[71]	CLOUDSIM	I	✔	Meta Heuristic	
29	IEEE	2019	[72]	CLOUDSIM	S	✔	Meta Heuristic	
30	IEEE	2021	[73]	NSFNET	S	✔	Machine learning	
31	ELSEVIER	2021	[74]	Python	S	✔	Non-Heuristic	✔
32	ELSEVIER	2021	[75]	NM	S	✔	Meta Heuristic	
33	WILEY	2021	[76]	Java	S	✔	Meta Heuristic	
34	IEEE	2019	[77]	CLOUD ANALYST	S	✔	Meta Heuristic	✔
35	SPRINGER	2019	[78]	C++	S	✔	Heuristic	
36	WILEY	2020	[79]	IFOGSIM	S	✔	Meta Heuristic	
37	ELSEVIER	2018	[80]	NETEM	S		Non-Heuristic	
38	IEEE	2019	[81]	PYTHON	S		No Algorithm	
39	ELSEVIER	2019	[82]	NM	S	✔	Heuristic	
40	SPRINGER	2017	[83]	IFOGSIM	S	✔	Heuristic	✔
41	ELSEVIER	2021	[84]	Docker	I	✔	Meta Heuristic	
42	ELSEVIER	2021	[85]	NS3	S	✔	Mathematical	
43	ELSEVIER	2022	[86]	IFOGSIM	S	✔	Meta Heuristic	
44	SPRINGER	2022	[87]	IFOGSIM	S	✔	Meta Heuristic	
45	T&F	2022	[88]	MATLAB	S	✔	Meta Heuristic	✔
46	IEEE	2019	[89]	JAVA	S	✔	Mathematical	
47	SPRINGER	2019	[90]	MATLAB	S	✔	Mathematical	

RQ4: What used tools for service and Energy management in fog computing?

Table 4 lists the simulation tools in the papers. In the following, we introduce the most common simulators based on their use. Two simulation tools, Cloud Sim and MATLAB have been applied in 11% of the research papers. Hence, they have the most popularity in the literature. After that, Python and ifogsim have also been used to simulate 9% of the research papers. A drawback of some papers is that they have not introduced their simulator.

According to our knowledge in this research, 34% of the papers do not mention their simulators.

Fig. 11 shows the percentage of use of different simulators in this research.

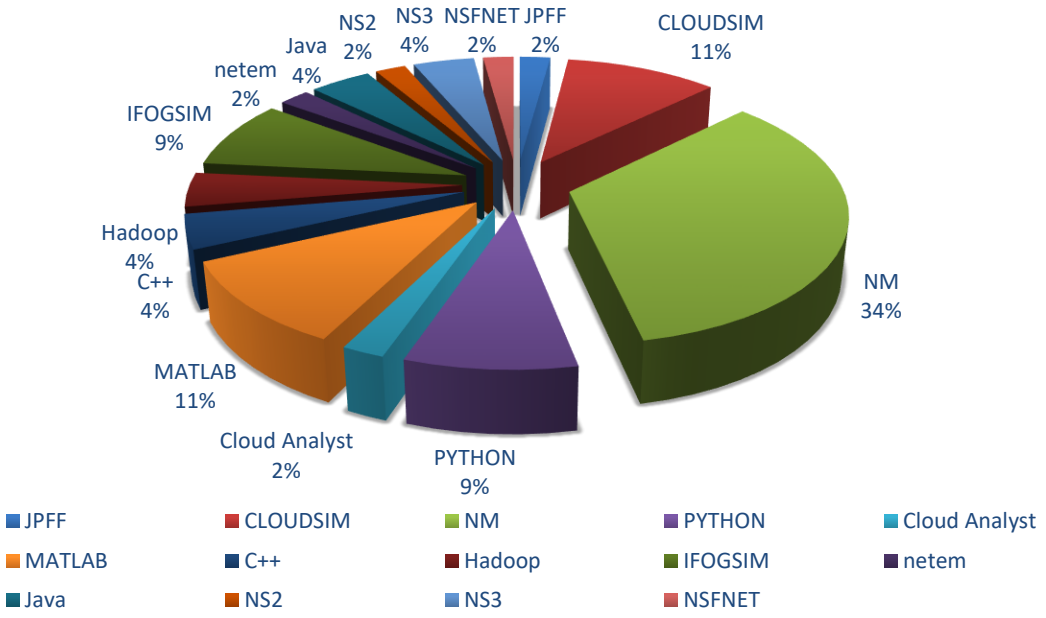


Fig. 11: The percentage of use of different simulators in the research literature.

RQ5: What is the significance of service and Energy management in fog computing?

The Fog architectures should allow storage, computing, and networking tasks to be dynamically migrated between the Tings, the Cloud, and the Fog. Therefore, the interfaces for Fog to interact with the Cloud, other Fogs, and the Things and users must:

1) Migration of the computing, control functions, and storage and facilitate flexibility through Other users
2) Permit appropriate access to the user for Fog services
3) Allow management of lifecycle in system and services effectively.

Task Scheduling will cause Energy consumption to be distributed among nodes in the network equally. In the real world, we are faced with limited resources. Resource allocation can increase efficiency and assign resources suitably to tasks.

RQ6: Which problems, future research directions, and challenges are identified concerning service and Energy management for future trends in fog computing?

In this paper, significant parts of fog computing have been investigated. In this section, future research directions for fog computing, existing gaps in studies of researchers, and open problems have been identified.

Security in Fog: Compared to the cloud, Fog has confronted the new security challenges. Fog is closer than the cloud, and its scattering is more than the cloud. Therefore, it is more vulnerable than the cloud and the centralized frameworks. The cloud works in a more secure structure, which is made by cloud operators, but the Fog works in a wide area. Security in fog computing is an open problem. Often, fog frameworks have more complexity than clouds. Therefore, they may not be able to manage their resources like clouds. In addition, the fog frameworks may not be able to recognize the dangers.

Not considering multi-objective: In some papers, authors only consider the efficiency of their method and do not examine other parameters. In fog computing, researchers must review not only cost and processing speed but also the parameters such as energy efficiency, delay, and throughput.

Not comparing with valid schemes: In various papers, an important challenge is that these schemes are not compared with valid methods. Often, they compare its method with basic methods, which is not appropriate in such a situation and shows a large difference. This means that the proposed methodology examines its efficiency compared to the classical schemes.

Simulation and implementation: The validity of a scheme is to implement it in the real environment. The simulation tool and its results are subject to real implementation. In general, some methods need the real implementation to evaluate their effectiveness. When a scheme implements in a real environment, it may encounter some problems; but, the simulation tool does not consider these constraints.

Scalability: In most simulations, the scalability of the network and the environment is limited. This causes some

problems, like mobility and delay. These problems may cause data loss in emergency issues or not sending data.

Heterogeneous resources and programs: Heterogeneous resources and programs are not considered important principles in many papers. Most schemes consider resources and virtual machines in a similar form. With this view, there are some problems, such as data distribution. These problems may produce different results for users. Some papers, which considered the heterogeneity of resources, have suggested strategies such as artificial intelligence and online learning. This problem can be expressed as an open issue.

Not considering the dynamics of the environment and real-time systems: Simulations are carried out as a fixed and predefined scenario. Geographic environments are usually defined in limited environments (even less than 1 km). Therefore, it is dangerous and unreliable that the proposed scheme is implemented in important applications such as healthcare, smart applications, etc. Real-time data processing in fog computing is not considered. Chaos theory and Lyapunov are provided to solve scheduling issues and online distribution of services/data, which can be used to solve future problems.

Being math of methods: The researchers formulate the fog computing problem as mixed-integer nonlinear programming and transform it into linear programming to increase its efficiency. It has some disadvantages and advantages. These schemes are not simple, and their details are not mentioned. They are also not implemented in a real environment. Therefore, these schemes are not confirmed. The main strength of these schemes is that they are optimized using linear and nonlinear functions.

Reliability: Fog computing can be considered due to safety mechanisms, compatibility of fog nodes, fault-tolerant, high-performance service availability, and other QoS parameters. Despite these parameters, an energy-efficient scheme, which reduces energy consumption, cannot provide network reliability. Energy harvest is inherently unreliable.

Architecture presentation: The contribution of architecture in research is extremely limited, and most architect are a particular model for their research. It is suggested to present a combined proposed architecture or a more efficient method for important medical issues requiring low communication and latency.

Using container-based methods: Container technology has emerged in cloud and fog in various studies and is even known as a container as a service, which helps in better allocating and placing resources in cloudy and fog environments. Very little research has been conducted or developed on container-based methods, and it seems that this method can have a change in energy reduction and latency.

Conclusion

In this study, SLR-based research was presented in the service/energy management approaches in the fog computing field. Service/energy management is a robust solution for improving energy efficiency. Based on existing research, this field has four domains in service management. In addition, the strengths and drawbacks of research should be considered. Challenges of each paper have been given to develop more efficient solutions in future research for service management. In this paper, we have presented an SLR method. In this paper, we have reviewed 1596 studies published from 2013 to 2022. Finally, we selected 47 studies focused on service/energy management approaches in the fog computing field. According to the RQ1, we deduced that Scheduling management has the highest use (i.e., 38%) in the field of service/energy management approaches in Fog. Based on RQ2, the statistical percentage of evaluation factors indicates that different QoS parameters in papers focus extremely on the cost (i.e., 49%). According to RQ3, it can be deduced that 81% of the research papers use the simulation environment to evaluate their proposed scheme in the fog computing platform. In addition, 15% of the papers utilize a real environment to implement their schemes in the fog computing environment. In the following, for answering the question, we deduced that some researchers use meta-heuristic and statistical methods in their schemes (about 36%). 30% of the papers have provided an architecture or framework along with their schemes. Based on RQ4, which is about simulation and modeling tools, we observed that Cloud Sim and Matlab are widely used as simulation tools in studies (about 11%). According to the SLR-based scheme, we cannot evaluate all existing studies. Therefore, some constraints are considered, including removing non-English papers, removing papers, which are less than 6 pages, conference papers, book chapters, and survey papers.

We believe that this research reviews the conceptual characteristics of service/energy management approaches in fog computing. In general, service management in a computational environment still needs further studies. It should be able to deal with its heterogeneity to improve energy efficiency by reducing requests. This research helps researchers and specialists to obtain a general understanding of this field and perform future research.

Author Contributions

This paper is the result of S.M. Hashemi Research project which is supervised and advised by A. Sahafi ,A.M. Rahmani, and M. Bohlouli respectively. S.M. Hashemi

wrote the manuscript. A. Sahafi and A.M. Rahmani sketched the research framework and the roadmap. Also, he analyzed the results and tabulated the outcome derived from excerpted literatures. In this line, M. Bohlouli searched in authentic journals to gather all relevant papers. they cooperatively summed up the work.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript.

Abbreviations

IOT	Internet of Things
FC	Fog computing
SLR	Systematic literature review
S	Simulation
I	Implementation
NM	Not Mentioned
ECCM	Energy-Efficient Cross-Layer-Sensing Clustering Method
DLR	Dynamic Thermal Rating
PLR	Packet Loss Rate
UWEE	User-Weighted Energy Efficiency
MINLP	Mixed-Integer Nonlinear Programming
ALM	Augmented LAGRANGE Method
MCPS	Medical Cyber-Physical Systems
POMDP	Partially Observable Markov Decision Process
MADS	Mesh Adaptive Direct Search
DBWA	Delay-Based Workload Allocation
DRQN	Deep Recurrent Q-Network
EONs	Elastic Optical Networks
EDTG	Energy-Efficient Deep Reinforced Traffic Grooming
IPSOLW	Improved Particle Swarm Optimization with Levy Walk

Reference

[1] M. H. Kashani, M. Madanipour, M. Nikravan, P. Asghari, E. Mahdipour, "A systematic review of IoT in healthcare: Applications, techniques, and trends," J. Network Comput. Appl., 192: 103164, 2021.

[2] T. Samizadeh Nikoui, A. M. Rahmani, A. Balador, H. Seyyed Javadi, "Internet of things architecture challenges: A systematic review," Int. J. Commun. Syst., 34: e4678, 2021.

[3] H. Kashani, A. M. Rahmani, N. Jafari Navimipour, "Quality of service-aware approaches in fog computing," Int. J. Commun. Syst., 33: e4340, 2020.

[4] A. Sadri, A. M. Rahmani, M. Saberikamarposhti, M. Hosseinzadeh, "Fog data management: A vision, challenges, and future directions," J. Network Comput. Appl., 174: 102882, 2021.

[5] M. Rahimi, N. Jafari Navimipour, M. Hosseinzadeh, M. H. Moattar, A. Darwesh, "Toward the efficient service selection approaches in cloud computing," Kybernetes, 51: 1388-1412, 2022.

[6] M. Sheikh Sofla, H. Kashani, E. Mahdipour, R. Faghih Mirzaee, "Towards effective offloading mechanisms in fog computing," Multimedia Tools Appl., 81: 1997-2042, 2022.

[7] F. Davami, S. Adabi, A. Rezaee, A. M. Rahmani, "Fog-based architecture for scheduling multiple workflows with high availability requirement," Computing, 104: 169-208, 2022.

[8] N. S. Soulegan, B. Barekatain, B. S. Neysiani, "MTC: minimizing time and cost of cloud task scheduling based on customers and providers needs using genetic algorithm," Int. J. Intell. Syst. Appl. (IJISA), 13: 38-51, 2021.

[9] A. Najafizadeh, A. Salajegheh, A. M. Rahmani, A. Sahafi, "Task scheduling in fog computing: a survey," J. Adv. Comput. Res., 11: 33-56, 2020.

[10] A. Najafizadeh, A. Salajegheh, A. M. Rahmani, A. Sahafi, "Multi-objective task scheduling in cloud-fog computing using goal programming approach," Cluster Comput., 25: 141-165, 2022.

[11] M. Abbasi, E. Mohammadi-Pasand, M. R. Khosravi, "Intelligent workload allocation in IoT–Fog–cloud architecture towards mobile edge computing," Comput. Commun., 169: 71-80, 2021.

[12] M. Laroui, B. Nour, H. Moungla, M. A. Cherif, H. Afifi, M. Guizani, "Edge and fog computing for IoT: A survey on current research activities & future directions," Comput. Commun., 180: 210-231, 2021.

[13] D. Wang, D. Zhong, A. Souri, "Energy management solutions in the internet of things applications: Technical analysis and new research directions," Cognit. Syst. Res., 67: 33-49, 2021.

[14] A. Mijuskovic, A. Chiumento, R. Bemthuis, A. Aldea, P. Havinga, "Resource management techniques for cloud/fog and edge computing: An evaluation framework and classification," Sensors, 21: 1832, 2021.

[15] B. Costa, J. Bachiega, L. Rebouças, A. P. F. Araujo, "Orchestration in fog computing: A comprehensive survey," ACM Computing Surveys (CSUR), 55(2): 1-34, 2022.

[16] M. Gill, D. Singh, "A comprehensive study of simulation frameworks and research directions in fog computing," Comput. Sci. Rev., 40: 100391, 2021.

[17] M. R. Alizadeh, V. Khajehvand, A. M. Rahmani, E. Akbari, "Task scheduling approaches in fog computing: A systematic review," Int. J. Commun. Syst., 33: e4583, 2020.

[18] M. Ghobaei-Arani, A. Souri, A. A. Rahmanian, "Resource management approaches in fog computing: a comprehensive review," J. Grid Comput., 18: 1-42, 2020.

[19] M. Mukherjee, L. Shu, D. Wang, "Survey of fog computing: Fundamental, network applications, and research challenges," IEEE Commun. Surv. Tutorials, 20: 1826-1857, 2018.

[20] A. Dasgupta, A. Gill, "Fog computing challenges: a systematic review," in Proc. ACIS 2017: 79, 2017.

[21] V. Pande, C. Marlecha, S. Kayte, "A review-fog computing and its role in the internet of things," Int. J. Eng. Res. Appl., 6: 2248-96227, 2016.

[22] P. Hu, S. Dhelim, H. Ning, T. Qiu, "Survey on fog computing: architecture, key technologies, applications and open issues," J. Network Comput. Appl., 98: 27-42, 2017.

[23] G. Caiza, M. Saeteros, W. Oñate, M. V. Garcia, "Fog computing at industrial level, architecture, latency, energy, and security: A review," Heliyon, 6: e03706, 2020.

[24] R. Zolfaghari, A. Sahafi, A. M. Rahmani, R. Rezaei, "Application of virtual machine consolidation in cloud computing systems," Sustainable Comput. Inf. Syst., 30: 100524, 2021.

[25] M. Effatparvar, M. Dehghan, A. M. Rahmani, "A comprehensive survey of energy-aware routing protocols in wireless body area sensor networks," J. Med. Syst., 40: 1-27, 2016.

[26] J. Ghomi, A. M. Rahmani, N. N. Qader, "Load-balancing algorithms in cloud computing: A survey," J. Network Comput. Appl., 88: 50-71, 2017.

[27] B. Kitchenham et al., "Systematic literature reviews in software engineering–a tertiary study," Information and software technology, vol. 52, pp. 792-805, 2010.

[28] A. Vakili, N. J. Navimipour, "Comprehensive and systematic review of the service composition mechanisms in the cloud environments," J. Network Comput. Appl., 81: 24-36, 2017.

[29] F. Tao, L. Zhang, Y. Liu, Y. Cheng, L. Wang, X. Xu, "Manufacturing service management in cloud manufacturing: Overview and future research directions," J. Manuf. Sci. Eng., 137(4): 040912, 2015.

[30] Y. Sun, Z. Xiao, D. Bao, J. Zhao, "An architecture model of management and monitoring on cloud services resources," in Proc. 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010.

[31] R. Mahmud, R. Kotagiri, R. Buyya, "Fog computing: A taxonomy, survey and future directions," Internet of Everything: Algorithms, Methodologies, Technologies and Perspectives, Springer, 2018.

[32] E. Casalicchio, L. Silvestri, "Architectures for autonomic service management in cloud-based systems," presented at the 2011 IEEE Symposium on Computers and Communications (ISCC), Kerkyra, Greece, 2011.

[33] B. Shojaiemehr, A. M. Rahmani, N. N. Qader, "Cloud computing service negotiation: A systematic review," Comput. Stand. Interfaces, 55: 196-206, 2018.

[34] E. Jafarnejad Ghomi, M. Rahmani, N. Qader, "Service load balancing, task scheduling and transportation optimisation in cloud manufacturing by applying queuing system," Enterp. Inf. Syst., 13: 865-894, 2019.

[35] L. Liu, D. Qi, N. Zhou, Y. Wu, "A task scheduling algorithm based on classification mining in fog computing environment," Wireless Commun. Mobile Comput., 2102348: 1-12, 2018.

[36] R. M. B. Abadi, A. M. Rahmani, S. H. Alizadeh, "Challenges of server consolidation in virtualized data centers and open research issues: a systematic literature review," J. Supercomput., 76: 2876-2927, 2020.

[37] S. Gu, Q. Zhuge, J. Yi, J. Hu, E. H. M. Sha, "Optimizing task and data assignment on multi-core systems with multi-port SPMs," IEEE Trans. Parallel Distrib. Syst., 26: 2549-2560, 2014.

[38] W. Lin, S. Xu, L. He, J. Li, "Multi-resource scheduling and power simulation for cloud computing," Inf. Sci., 397: 168-186, 2017.

[39] M. A. A. Faruque, F. Ahourai, "A model-based design of cyber-physical energy systems," presented at the 2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC), Singapore, 2014.

[40] F. Faruque, K. Vatanparvar, "Energy management-as-a-service over fog computing platform," IEEE Internet Things J., 3: 161-169, 2015.

[41] H. Barati, A. Movaghar, A. M. Rahmani, "EACHP: Energy aware clustering hierarchy protocol for large scale wireless sensor networks," Wireless Pers. Commun., 85: 765-789, 2015.

[42] S. E. Dashti, A. M. Rahmani, "Dynamic VMs placement for energy efficiency by PSO in cloud computing," J. Exp. Theor. Artif. Intell., 28: 97-112, 2016.

[43] K. Chronaki et al., "Task scheduling techniques for asymmetric multi-core systems," IEEE Trans. Parallel Distrib. Syst., 28: 2074-2087, 2016.

[44] Z. Sun et al., "An energy-efficient cross-layer-sensing clustering method based on intelligent fog computing in WSNs," IEEE Access, 7: 144165-144177, 2019.

[45] N. Kaur, S. K. Sood, "An energy-efficient architecture for the Internet of Things (IoT)," IEEE Syst. J., 11: 796-805, 2015.

[46] A. Rafiq, W. Ping, W. Min, M. Saleh, "Fog assisted 6TiSCH tri-layer network architecture for adaptive scheduling and energy-efficient offloading using rank-based Q-learning in smart industries," IEEE Sensors J., 21: 25489-25507, 2021.

[47] M. Dabbaghjamanesh, A. Kavousi-Fard, Z. Y. Dong, "A novel distributed cloud-fog based framework for energy management of networked microgrids," IEEE Trans. Power Syst., 35: 2847-2862, 2020.

[48] G. Rathee, R. Sandhu, H. Saini, M. Sivaram, V. Dhasarathan, "A trust computed framework for IoT devices and fog computing environment," Wireless Networks, 26: 2339-2351, 2020.

[49] Z. Ren, T. Lu, X. Wang, W. Guo, G. Liu, S. Chang, "Resource scheduling for delay-sensitive application in three-layer fog-to-cloud architecture," Peer-to-Peer Networking Appl., 13: 1474-1485, 2020.

[50] S. K. Mani, I. Meenakshisundaram, "Improving quality-of-service in fog computing through efficient resource allocation," Comput. Intell., 36: 1527-1547, 2020.

[51] X. Chen, Y. Zhou, L. Yang, L. Lv, "User satisfaction oriented resource allocation for fog computing: A mixed-task paradigm," IEEE Trans. Commun., 68: 6470-6482, 2020.

[52] L. Ni, J. Zhang, C. Jiang, C. Yan, K. Yu, "Resource allocation strategy in fog computing based on priced timed petri nets," IEEE Internet Things J., 4: 1216-1228, 2017.

[53] L. Gu, D. Zeng, S. Guo, A. Barnawi, Y. Xiang, "Cost efficient resource management in fog computing supported medical cyber-physical system," IEEE Trans. Emerging Top. Comput., 5: 108-119, 2015.

[54] R. Basir, S. Qaisar, M. Ali, M. Naeem, A. Anpalagan, "Energy efficient resource allocation in cache-enabled fog networks," Trans. Emerging Telecommun. Technol., 32: e4343, 2021.

[55] S. Lu, J. Wu, Y. Duan, N. Wang, J. Fang, "Towards cost-efficient resource provisioning with multiple mobile users in fog computing," J. Parallel Distrib. Computing, 146: 96-106, 2020.

[56] Y. Sun, N. Zhang, "A resource-sharing model based on a repeated game in fog computing," Saudi J. Biol. Sci., 24: 687-694, 2017.

[57] S. Kalantary, J. Akbari Torkestani, A. Shahidinejad, "Resource discovery in the internet of things integrated with fog computing using Markov learning model," J. Supercomput., 77: 13806-13827, 2021.

[58] S. Rodrigo, N. Ls, "Location of fog nodes for reduction of energy consumption of end-user devices," IEEE Trans. Green Commun. Networking, 4: 593-605, 2020.

[59] M. Guo, L. Li, Q. Guan, "Energy-efficient and delay-guaranteed workload allocation in IoT-edge-cloud computing systems," IEEE Access, 7: 78685-78697, 2019.

[60] N. Mazumdar, A. Nag, J. P. Singh, "Trust-based load-offloading protocol to reduce service delays in fog-computing-empowered IoT," Comput. Electr. Eng., 93: 107223, 2021.

[61] D. Tychalas, H. Karatza, "A scheduling algorithm for a fog computing system with bag-of-tasks jobs: Simulation and performance evaluation," Simul. Modell. Pract. Theory, 98: 101982, 2020.

[62] S. Bitam, S. Zeadally, A. Mellouk, "Fog computing job scheduling optimization based on bees swarm," Enterp. Inf. Syst., 12: 373-397, 2018.

[63] R. O. Aburukba, M. AliKarrar, T. Landolsi, K. El-Fakih, "Scheduling internet of things requests to minimize latency in hybrid Fog–Cloud computing," Future Gener. Comput. Syst., 111: 539-551, 2020.

[64] S. Chouikhi, M. Esseghir, L. Merghem-Boulahia, "Energy consumption scheduling as a fog computing service in smart grid," IEEE Trans. Serv. Comput., 16: 1144-1157, 2022.

[65] R. Xie, Q. Tang, C. Liang, F. R. Yu, T. Huang, "Dynamic computation offloading in IoT fog systems with imperfect channel-state information: A POMDP approach," IEEE Internet Things J., 8: 345-356, 2020.

[66] X. Li, Z. Zang, F. Shen, Y. Sun, "Task offloading scheme based on improved contract net protocol and beetle antennae search algorithm in fog computing networks," Mobile Networks Appl., 25: 2517-2526, 2020.

[67] H. Mahini, A. M. Rahmani, S. M. Mousavirad, "An evolutionary game approach to IoT task offloading in fog-cloud computing," J. Supercomput., 77: 5398-5425, 2021.

[68] D. Wang, Z. Liu, X. Wang, Y. Lan, "Mobility-aware task offloading and migration schemes in fog computing networks," IEEE Access, 7: 43356-43368, 2019.

[69] C. Ren, X. Lyu, W. Ni, H. Tian, and R. P. Liu, "Distributed online learning of fog computing under nonuniform device cardinality," IEEE Internet Things J., 6: 1147-1159, 2018.

[70] O. K. Shahryari, H. Pedram, V. Khajehvand, M. D. TakhtFooladi, "Energy and task completion time trade-off for task offloading in fog-enabled IoT networks," Pervasive Mob. Comput., 74: 101395, 2021.

[71] M. Yang, H. Ma, S. Wei, Y. Zeng, Y. Chen, Y. Hu, "A multi-objective task scheduling method for fog computing in cyber-physical-social services," IEEE Access, 8: 65085-65095, 2020.

[72] J. Xu, Z. Hao, R. Zhang, X. Sun, "A method based on the combination of laxity and ant colony system for cloud-fog task scheduling," IEEE Access, 7: 116218-116226, 2019.

[73] R. Zhu, S. Li, P. Wang, M. Xu, S. Yu, "Energy-efficient deep reinforced traffic grooming in elastic optical networks for cloud–fog computing," IEEE Internet Things J., 8: 12410-12421, 2021.

[74] M. Hussain, L. F. Wei, A. Lakhan, S. Wali, S. Ali, A. Hussain, "Energy and performance-efficient task scheduling in heterogeneous virtualized cloud computing," Sustainable Comput. Inf. Syst., vol. 30: 100517, 2021.

[75] M. Abd Elaziz, L. Abualigah, I. Attiya, "Advanced optimization technique for scheduling IoT tasks in cloud-fog computing environments," Future Gener. Comput. Syst., 124: 142-154, 2021.

[76] M. Abdel-Basset, R. Mohamed, R. K. Chakrabortty, M. J. Ryan, "IEGA: an improved elitism-based genetic algorithm for task scheduling problem in fog computing," Int. J. Intell. Syst., 36: 4592-4631, 2021.

[77] Z. A. Khan et al., "Energy management in smart sectors using fog-based environment and meta-heuristic algorithms," IEEE Access, 7: 157254-157267, 2019.

[78] L. Stavrinides, H. D. Karatza, "A hybrid approach to scheduling real-time IoT workflows in fog and cloud environments," Multimedia Tools Appl., 78: 24639-24655, 2019.

[79] M. Ayoubi, M. Ramezanpour, R. Khorsand, "An autonomous IoT service placement methodology in fog computing," Softw: Pract. Exper., 51: 1097-1120, 2021.

[80] V. B. Souza et al., "Towards a proper service placement in combined Fog-to-Cloud (F2C) architectures," Future Gener. Comput. Syst., 87: 1-15, 2018.

[81] J. Li et al., "Service migration in fog computing enabled cellular networks to support real-time vehicular communications," IEEE Access, 7: 13704-13714, 2019.

[82] D. Zhao, G. Sun, D. Liao, S. Xu, V. Chang, "Mobile-aware service function chain migration in cloud–fog computing," Future Gener. Comput. Syst., 96: 591-604, 2019.

[83] O. Skarlat, M. Nardelli, S. Schulte, M. Borkowski, P. Leitner, "Optimized IoT service placement in the fog," Serv. Oriented Comput. Appl., 11: 427-443, 2017.

[84] B. V. Natesha, G. Ram, "Adopting elitism-based Genetic Algorithm for minimizing multi-objective problems of IoT service placement in fog computing environment," J. Network Comput. Appl., 178: 102972, 2021.

[85] S. O. Ogundoyin, I. A. Kamil, "A trust management system for fog computing services," Internet of Things, 14: 100382, 2021.

[86] M. Ghobaei-Arani, A. Shahidinejad, "A cost-efficient IoT service placement approach using whale optimization algorithm in fog computing environment," Expert Syst. Appl., 200: 117012, 2022.

[87] N. Sarrafzade, R. Entezari-Maleki, L. Sousa, "A genetic-based approach for service placement in fog computing," J. Supercomput., 78: 10854-10875, 2022.

[88] M. Salimian, M. Ghobaei-Arani, A. Shahidinejad, "An evolutionary multi-objective optimization technique to deploy the IoT Services in fog-enabled Networks: an autonomous approach," Appl. Artif. Intell., 36: 2008149, 2022.

[89] A. Yousefpour et al., "Fogplan: A lightweight qos-aware dynamic fog service provisioning framework," IEEE Internet Things J., 6: 5080-5096, 2019.

[90] Y. Chen, Y. Wang, D. Gong, "Fog computing support scheme based on fusion of location service and privacy preservation for QoS enhancement," Peer-to-Peer Networking Appl., 12: 1480-1488, 2019.

Biographies

Sayed Mohsen Hashemi received the bachelor's degree in information technology from Payame Noor University, Iran, in 2011, and the master's degree in software engineering from Islamic Azad University, Meybod Branch, Iran, in 2013. He is currently pursuing the Ph.D. degree in computer engineering with Islamic Azad University, Qeshm, Iran. His research interests include cloud computing, fog computing, the Internet of Things, routing algorithm in computer networks, data mining, machine learning, and meta heuristic algorithms.

- Email: mohsenpnu2009@gmail.com
- ORCID: 0000-0002-6506-9987
- Web of Science Researcher ID: HGU-0399-2022
- Scopus Author ID: NA
- Homepage: NA

Amir Sahafi received the B.Sc. degree in computer engineering from Shahed University, Tehran, Iran, in 2005, and the M.Sc. and Ph.D. degrees in computer engineering from the Science and Research Branch, Islamic Azad University, Tehran, in 2007 and 2012, respectively. He is currently an Assistant Professor with the Department of Computer

Engineering, South Tehran Branch, Islamic Azad University, Tehran. His current research interest includes distributed and cloud computing.

- Email: sahafi@iau.ac.ir
- ORCID: 0000-0002-6555-670X
- Web of Science Researcher ID: NA
- Scopus Author ID: 24528878600
- Homepage: https://stb.iau.ir/faculty/a-sahafi/fa

Amir Masoud Rahmani received the B.S. degree in computer engineering from Amir Kabir University, Tehran, in 1996, the M.S. degree in computer engineering from the Sharif University of Technology, Tehran, in 1998, and the Ph.D. degree in computer engineering from IAU University, Tehran, in 2005. He is currently a Professor of computer engineering. His research interests include distributed systems, the Internet of Things, and evolutionary computing.

- Email: rahmani@srbiau.ac.ir
- ORCID: 0000-0001-8641-6119
- Web of Science Researcher ID: K-2702-2013
- Scopus Author ID: 57204588830
- Homepage: https://srb.iau.ir/faculty/a-rahmani/fa

Mahdi Bohlouli received the Ph.D. degree from the University of Siegen, with the main focus on statistical regeneration and scalable clustering of big data using Map Reduce in the Hadoop ecosystem. He was the Group Leader of web search and data mining at the Institute for Web Science and Technologies (WeST), University of Koblenz, Germany, and a Senior Research Associate and a Project Manager at the Institute of Knowledge-Based Systems (KBS), University of Siegen, Germany. He is currently an Assistant Professor of data science and machine learning at the Institute for Advanced Studies in Basic Sciences (IASBS) with the main focus on the large-scale data analysis and next generation AI, in particular reinforcement, generative and self-attentive learning algorithms. He leads the Intelligent Systems Group (ISG), IASBS, as well as various workshops and conferences, being co-located with ICML, WWW, IEEE Big Data, and Semantics conferences and many more. He has promising scholar records in various AI fields. He was also involved in leading up various industrial software projects in the IT sector.

- Email: me@bohlouli.com
- ORCID: 0000-0002-6659-5524
- Web of Science Researcher ID:
- Scopus Author ID: 55809185800
- Homepage: www.bohlouli.com

How to cite this paper:
S. M. Hashemi, A. Sahafi, A. M. Rahmani, M. Bohlouli, "Service and energy management in fog computing: A taxonomy approaches, and future directions," J. Electr. Comput. Eng. Innovations, 12(1): 15-38, 2024.

DOI: 10.22061/jecei.2023.9482.624

URL: https://jecei.sru.ac.ir/article_1886.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research Paper

Application of Grey Wolf Optimization Algorithm with Aggregation Function on Designing Interleaved Boost Converter

S. M. Naji Esfahani [1], *S. H. Zahiri* [1,*], *M. Delshad* [2]

[1]*Department of Electrical Engineering, Faculty of Engineering, University of Birjand, Birjand, Iran.*

[2]*Department of Electrical Engineering, Islamic Azad University Isfahan (Khorasgan) Branch, Isfahan, Iran.*

Article Info

Article History:
Received 11 Marsh 2023
Reviewed 21 April 2023
Revised 17 May 2023
Accepted 14 June 2023

Keywords:
Interleaved boost converter
Non-Minimum phase system
Optimized proportional integral controller
Grey wolf optimization algorithm with aggregation function definition (GWO_AF)
Switch-Mode power supply

[*]Corresponding Author's Email Address: *hzahiri@birjand.ac.ir*

Abstract

Background and Objectives: The interleaved approach has a long history of use in power electronics, particularly for high-power systems. The voltage and current stress in these applications exceed the tolerance limit of a power element. The present paper introduces an improved version of an interleaved boost converter, which uses voltage mode control. The objectives of this research are improvement in the interleaved boost converter's performance in terms of the temporal parameters associated with settling duration, rising time, and overshoot.

Methods: An improved PI controller (proportional integral controller) is used for adjusting the proposed converter's output voltage. In the present work, the Grey Wolf Optimization algorithm with aggregation function definition (GWO_AF) is utilized to adjust the free coefficients of the PI controller. The closed-loop dynamic performance and stability can be improved by designing and implementing an optimized PI controller.

Results: The improvement of the freedom degree in the interleaved boost converter resulted from the existence of a few power switches in a parallel channel in the proposed circuit. An additional advantage of the interleaved boost converter, compared to the conventional one, is that it produces a lower output voltage ripple.

Conclusion: The usage of multi-objective optimization algorithms in designing a PI controller can significantly improve the performance parameters of an interleaved boost converter. Also, our findings indicated the excellent stability of the proposed converter when connected to the network.

Introduction

The interleaved boost converters have a wide range of usage in industry and high-power electronics systems. The voltage and current stress in these applications exceed the tolerance limit of a power element. Utilizing numerous five sources of power in parallel or series is one approach to solve this issue. This approach will result in voltage and current split in parallelization of the power elements. Parallelization of the power converters is another method that deserves mention for addressing this issue [1], [2]. If there is a proper control for the converter and several boost converters in parallel, it can improve the performance of the interleaved boost converter. Controlling the switch on each boost converter is performed using the interleaved pulse gates [3], [4]. The switching frequency is closely related to the frequency of the gate pulses. However, the phases of the control signals can be changed. The input current is across the switches of the interleaved boost converter, all of which are connected in parallel. As a result, the converter significantly outperforms the conventional boost converter in terms of reliability and effectiveness. In the

interleaved boost converter, the freedom degree is enhanced because of the existence of a few power switches in a parallel path. This feature leads to the enhancement of numerous crucial characteristics, such as the elimination of harmonics, enhanced effectiveness, reduced conductive loss, improved power density, and line tolerance [5].

In addition, the interleaved boost converter yields a lower output voltage ripple. As a result, the output filter's size and losses can considerably decrease compared to the ordinary boost converter. The control signals in the interleaved method are separated typically based on the similarity of the switching frequencies. Thus, the waveforms resulting from input and output currents are associated with lower ripples and fewer harmonics compared to the ordinary boost converter. The elimination of the low-frequency harmonics leads to a considerable reduction in the switching and conductive losses and the electromagnetic interface surfaces in the interleaved boost converter.

A two-phase converter includes two output stages with a 180o phase difference. If the current is divided into two paths, the power of the conductive loss will decrease, resulting in an increase in the total efficiency compared to the single-phase converter. The combination of two phases in the output 35 capacitor leads to doubling the effective frequency and reducing the voltage ripple. Similarly, the pulses of the power derived from the input capacitor are alternated, causing a reduced current ripple. For achieving the specific requirement in high-power applications, it is better to use the interleaved multi-channel converter, especially in cases where there are power pieces with limited function. To give practical examples of such cases, we can mention designing the switch-mode power supplies, power modules of spaceships, automobile engineering, hybrid electrical vehicles, and satellite applications [6]-[8]. For the use of the photovoltaic solar panels' maximum power, the best solution is to utilize the interleaved boost converters. In such a case, it is necessary to consider its rapid transient response and the absence of high-frequency ripple since they might interfere with the SPV system. These converters can be highly useful for those applications in which there is a demand for a low ripple or a very high tolerance by the consumer.

So far, several studies have been conducted on the two-phase interleaved boost converters published in scientific journals. Increasing the number of phases in the interleaved boost converter makes the circuit design more complicated.

Thus, in the present work, for simplicity, the two-phase interleaved boost converter has been assumed in the continuous conduction mode (CCM) [9]. In recent years, many changes have been made to the technology of DC-DC switching converters. The switch-mode power converters require a rapid transient response to supply the new-generation modern microprocessors, electrical vehicles, DPSs, and the integration of renewable energy resources in a grid.

Therefore, in closed-loop converters, the transient performance is a major factor in designing the power supplies in practice. However, there is a pole on the right side of the panel in some of the converters, for instance, interleaved boost converters and boost converters. This pole exists in the function of control-to-output transfer for the CCM performance [10]. Thus, an initial undershoot has been observed for the step input in these converters, which exhibit a poor dynamic performance attributed to the non-minimum phase. These unwanted undershoots have a higher importance in the cases where the RHP zero approaches the origin. Such an RHP zero restricts the closed-loops bandwidth and yields a slower dynamic for the converter [11]-[17].

Many researchers have attempted for several years to improve the power performance of the converters based on improving their efficiency, decreasing the losses, and removing the noise. However, the old methods failed to provide the desired results.

This prompted the researchers to search for new techniques and modify and improve the old ones. Improving the control performance could lead to some success. Because of their simplicity, the validity of the design method, and well-known behaviors, PI controllers are widely used in industries. Nevertheless, the commonly used PI controllers cannot fully solve the non-minimum phase problem [18]. The current paper evaluates the classic 80 PI controller's performance for a boost interleaved converter. For this purpose, the controller parameter regulation method, accompanied by the GWO_AF algorithm, aims to achieve the proposed closed-loop performance by modifying the reference voltage and comparing it with the PI controller. This process led to the stability of the converter [19]-[23].

Several physical mechanisms inspire the algorithms for designing the optimized controllers, some of which include the biological molecular behavior and characteristics of the insect's swarm and photo and biological systems.

So far, no specific algorithm has been presented to obtain the optimal solution for the problems related to the optimization issue [24]. Yet, some of the algorithms may yield better solutions compared to others. As demonstrated in Fig. 1, we employed the controller feedback loop, which measures the output voltage V_{out} and its comparison with the desired reference voltage V_{ref}.

$E^{(t)}$ error is obtained by measuring the converter output voltage and comparing V_{out} with the reference

voltage responsible for error minimization control by generating the desired control signal U(t).

If there is no use of optimization algorithms for choosing optimized parameters of K_p and K_i, overshoot, rise time error, settling time error, and steady-state error will happen to the converter output voltage. This controller mitigates the mentioned errors in a wide range of variations and differences in the input, output, and reference. In this research paper, the GWO_AF method has been employed for designing the PI controller of the two-phase boost interleaved converter [25], [26]. The GWO_AF algorithm is based on the swarm intelligence optimization methods, simply implemented [27]-[29], and has higher efficiency compared to other optimization methods of the converter controllers.

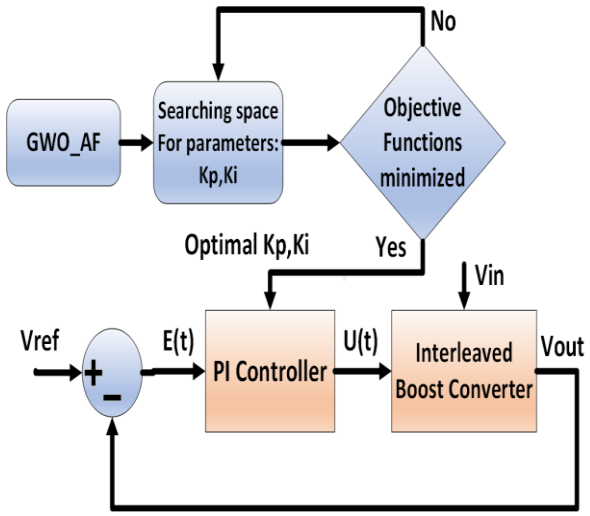


Fig. 1: Controller system process schematic.

In this study, we introduced the optimized GWO_AF-based PI controller for the boost interleaved converter. In the proposed approach, obtaining the optimized K_p and Ki is followed by comparing four different objective functions, integral absolute error (IAE), integral square error (ISE), integral time square error (ITSE), and integral time absolute error (ITAE) to find the best performance of the cost functions. The simulation results have been given in this paper to demonstrate the effect of the optimized controller on the proposed controller.

The rest of the paper is organized as follows: The small-signal analysis of two-phase interleaved boost converters is described in section 2.

Section 3 is devoted to the controller system design. The evaluation method and forming of the objective functions are stated in section 4. The calculative implementation of GWO_AF is covered in section 5. Section 6 presents the simulation results and investigation. Section 7 concludes the paper and provides future trends. Finally, section 8 is for compliance with Ethical Standards.

Description of the Two-Phase Interleaved Boost Converters

Fig. 2 shows a schematic view of the two-phase interleaved boost converter. This schematic image includes the inductor L_1 connected to L_2 in parallel, the switch Q_1 connected to Q_2 in parallel, and the diode D_1 connected to D_2 in parallel. On this basis, there are two input and output circuits in parallel. All similar pieces have been designed for better performance of the interleaved converter. The switching operation by the gate signals of the two switches is such that when one switch is in the maximum state, the other is in the minimum state with a phase difference of 180 degrees [28].

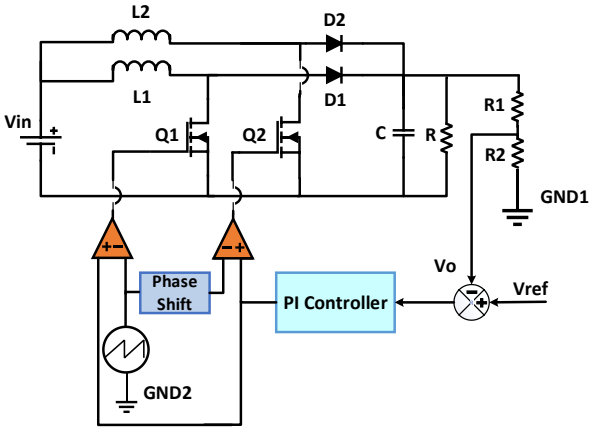


Fig. 2: Schematic diagram of the interleaved boost circuit with two-phase control.

If we assume that the inductor's current ripple is equal to 20% of the inductor's average current, we can obtain the value of the inductor from (1). D_{max} is the maximum value of the duty coefficient, which is 0.75, and V_{min} is the minimum value of the input voltage (48V).

$$L_{phase} = \frac{V_{in} \times D}{f_S \delta_{il}} \quad (1)$$

Considering 2% of the peak-to-peak capacitor ripple, the value of the capacitor can be obtained using (2).

$$\delta V_{out} = \frac{V_{in} \times D}{T_S C_{out} R} \quad (2)$$

Operation Modes

The state space averaging technique is used for analyzing the interleaved boost converter. By applying this mathematical model, the converter's switching function is defined in four switching methods. The state equations obtained for the converter's operation mode of Fig. 3 are as follows [28].

In Mode 1, Switches Q_1 and Q_2 are in the on mode, but diodes D_1 and D_2 are in the Off mode. Fig. 5 shows the

equivalent circuit for this mode. The following formulas define the function of Mode 1, in which the current of Inductor i_{Li} is assumed as the state variable, and the voltage of Capacitor V0 is assumed as the third mode variable.

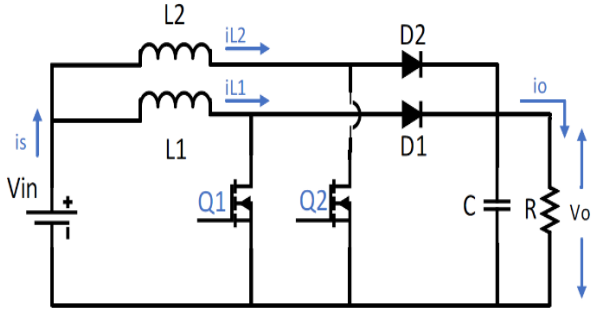


Fig. 3: Schematic diagram of the interleaved boost circuit.

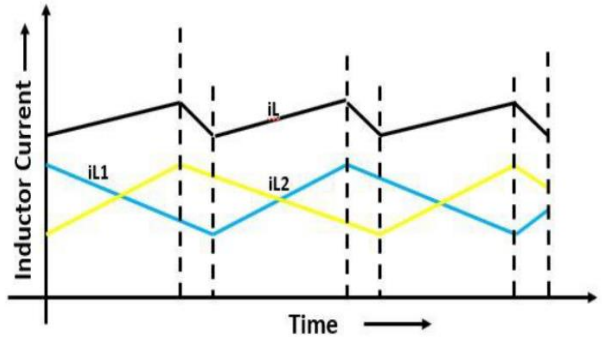


Fig. 4: Ideal waveform for the interleaved boost converter.

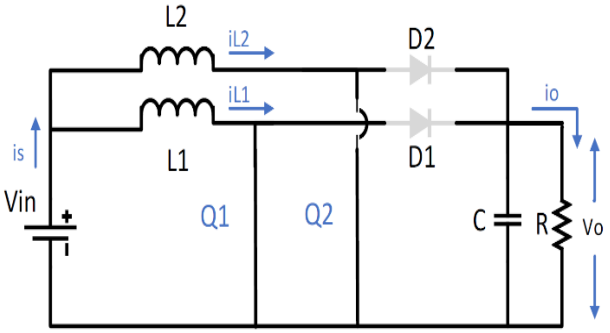


Fig. 5: Equivalent circuit for Mode 1.

$$\frac{di_{L1}}{dt}=\frac{V_s}{L_1} \tag{3}$$

$$\frac{d_{vo}}{dt}=\frac{V_O}{RC} \tag{4}$$

$$\frac{di_{L2}}{dt}=\frac{V_s}{L_2} \tag{5}$$

$$A=\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & \frac{-1}{RC}\end{bmatrix} B_1=\begin{bmatrix}\frac{1}{L_1}\\ \frac{1}{L_1}\\ V_O\end{bmatrix} \tag{6}$$

In Mode 2, Switches Q_1 and Q_2 are in the On and Off modes, respectively. Also, Diodes D_1 and D_2 are in the Off and on modes, respectively. Mode 2 is shown in Fig. 6.

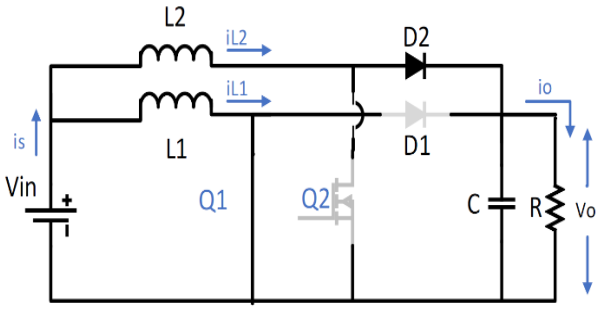


Fig. 6: Equivalent circuit for Mode 2.

$$\frac{di_{L1}}{dt}=\frac{V_s}{L_1} \tag{7}$$

$$\frac{di_{L2}}{dt}=\frac{V_s}{L_2}-\frac{V_0}{L_2} \tag{8}$$

$$\frac{d_{vo}}{dt}=\frac{i}{L_2}-\frac{V_0}{RC} \tag{9}$$

$$A=\begin{bmatrix}0 & 0 & \frac{1}{L_1}\\ 0 & 0 & 0\\ 0 & \frac{1}{C} & \frac{-1}{RC}\end{bmatrix} B_1=\begin{bmatrix}\frac{1}{L_1}\\ \frac{1}{L_1}\\ 0\end{bmatrix} \tag{10}$$

In Mode 3, Switch Q_1 is in the off mode, but Switch Q_2 is in the on mode. Also, Diodes D_1 and D_2 are in the on and off modes, respectively. Fig. 7 shows the performance of the interleaved boost converter in Mode 3.

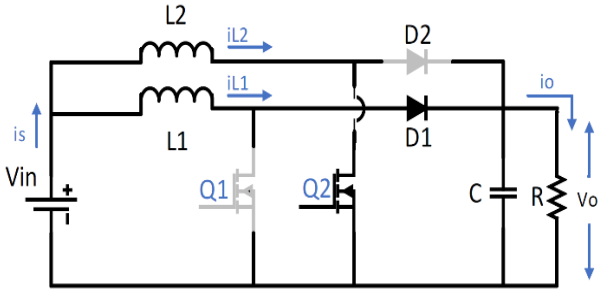


Fig. 7: Equivalent circuit for Mode 3.

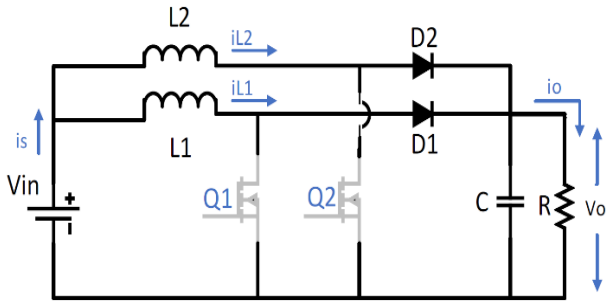


Fig. 8: Equivalent circuit for Mode 4.

$$\frac{di_{L1}}{dt} = \frac{V_S}{L_1} - \frac{V_O}{L_1} \tag{11}$$

$$\frac{di_{L2}}{dt} = \frac{V_S}{L_2} \tag{12}$$

$$\frac{dv_O}{dt} = \frac{i_{L1}}{C} - \frac{V_O}{RC} \tag{13}$$

$$A_3 = \begin{bmatrix} 0 & 0 & \frac{-1}{L_1} \\ 0 & 0 & 0 \\ \frac{1}{C} & 0 & \frac{-1}{RC} \end{bmatrix} B_3 = \begin{bmatrix} \frac{1}{L_1} \\ \frac{1}{L_2} \\ 0 \end{bmatrix} \tag{14}$$

In Mode 4, Switches Q_1 and Q_2 are in the Off mode, but Diodes D_1 and D_2 are in the on mode. Fig. 8 displays the equivalent circuit for this mode.

$$\frac{di_{L1}}{dt} = \frac{V_S}{L_1} - \frac{V_O}{L_1} \tag{15}$$

$$\frac{di_{L2}}{dt} = \frac{V_S}{L_2} - \frac{V_O}{L_2} \tag{16}$$

$$\frac{dv_O}{dt} = \frac{i_{L1}}{C} + \frac{i_{L2}}{C} - \frac{V_O}{RC} \tag{17}$$

$$A = \begin{bmatrix} 0 & 0 & \frac{1}{L_1} \\ 0 & 0 & \frac{1}{L_2} \\ 0 & \frac{1}{C} & \frac{-1}{RC} \end{bmatrix} B_4 = \begin{bmatrix} \frac{1}{L_1} \\ \frac{1}{L_2} \\ 0 \end{bmatrix} \tag{18}$$

The state equations and matrix coefficients for the interleaved boost converter are as follows:

$$X = AX + BU \tag{19}$$

$$Y = CX + DU \tag{20}$$

$$[A] = A_1 d_1 + A_2 d_2 + A_3 d_3 + A_4 d_4 \tag{21}$$

$$[B] = B_1 d_1 + B_2 d_2 + B_3 d_3 + B_4 d_4 \tag{22}$$

$$D = d_1 + d_2 + d_3 + d_4 \tag{23}$$

Using the SSA method, the interleaved boost converter's output control transfer function in its final form is obtained as follows. do is the duty coefficient, and N is the number of phases in the interleaved converter. It is evident that there is an RHP zero in the interleaved boost converter's transfer function [30]. By substituting the converter's parameters from Table 1, the interleaved boost converter's transfer function is found, which is as follows: The transfer function of the interleaved boost converter is as follows:

$$\frac{V_{O(S)}}{V_{in(S)}} = \frac{1 + sRC}{(1-D)\left[1 + s\frac{L}{R(1-D)^2} + s^2\frac{LC}{(1-D)^2}\right]} \tag{24}$$

$$\frac{V_{in(S)}}{d_{(S)}} = \frac{1 + sRC}{(1-D)\left[1 + s\frac{L}{R(1-D)^2} + s^2\frac{LC}{(1-D)^2}\right]} \tag{25}$$

$$T_{P-IBC(S)} = \frac{V_{O(S)}}{d_{(S)}} = \frac{2NV_{in}}{\left(r + 2Nd^2R_L\right)C} * \frac{-s + (2Nd^2R_L - r)/L_C}{(s + \frac{r}{L_C})(s + \frac{1}{R_L C}) + \frac{2Nd^2}{L_C C}} \tag{26}$$

$$T_{P-IBC(S)} = \frac{V_{O(S)}}{d_{(S)}} = \frac{-1.666^5 s + 7.675^9}{s^2 + 328.5s + 1.441^7} \tag{27}$$

Table 1: The interleaved boost converter design values

No.	Description	Design parameter value
1	Input voltage	48 V
2	Output voltage	240 V
3	Output power	54.8 W
4	Load current	0.23 A
5	Switching frequency	20 KHz
6	L_1 and L_2 inductance	1 mH
7	Capacity	1 µF
8	Load resistance	3200 Ω

Designing the Controller

For achieving the desired closed-loop performance of the converter necessitates the controller's presence. Besides, the controller can facilitate the circuit transfer function's formation for obtaining the system's general stability and the rapid transient response. There are several kinds of analog compensators, namely the RC network and amplifier.

The forward controller, which is also known as the (PD) controller, is commonly used for the phase margin enhancement in a bipolar system (two-pole system). This type of controller exhibits sensitivity to the noise resulting from the derivative function in it. The use of reverse PI controller (proportional-integral controller) causes increase in the low-frequency gain in loops, leading to stronger output at DC and the frequencies that are lower than the frequency of loop crossover [31], [32].

There are various types of classic controllers, namely forward-backward, PI, PID, etc. These controllers have been developed aiming to make sure of the converters' desirable performance under special conditions. However, for the CCM performance in the interleaved boost controller, there exists an RHP zero in the output control transfer function.

Thus, the non-minimal phase problem results in the poor dynamic effectiveness of the interleaved boost converter. Because of the RHP zero, the closed-loop system's bandwidth is restricted and the converter's dynamics is decreased. Hence, exhibiting a good response to the uncertainty of parameters, line, and instant load changes would be very difficult for the ordinary PID controller. According to the reports, the optimal PI controllers, which are based on the metaheuristic algorithms, can enhance the DC-DC converters' dynamic performance [33], [34].

Designing the GWO_AF-Based Optimal PI Controller

It sould be mentioned that many of other hyper heuristic algorithms can be employed for designing PI cotroller. In this paper GWO algorithm has been chosen due to its good performance in this research area [36]-[41].

As can be seen in Fig. 9, the converter's output voltage is compared to the reference voltage, and the generated error signal passes through the optimal classic controller, the coefficients of which are obtained by using the GWO_AF algorithm. Subsequently, to generate the pulse width modulation (PWM) signal, the created control signal is compared to the high-frequency triangle waveform [32].

This PWM signal, after passing the Astable multi-vibrator and optoisolator, produces the signal of the MOSFET switches Q_1 and Q_2.

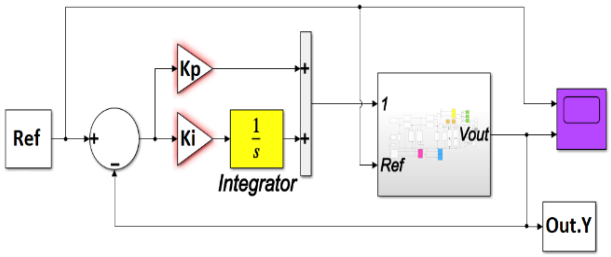


Fig. 9: Circuit of PI controller.

The optimal parameters of the PI controller are obtained using the GWO_AF algorithm considering the best performance of the fitness function to optimize the overshoot (M_p), the rise time (T_r), and the settling time (Ts).

Fig. 10 shows the diagram of adjusting the coefficients of the PI controller by the GWO_AF algorithm.

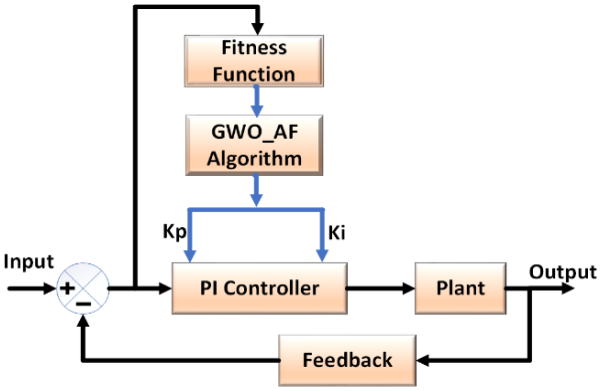


Fig. 10: Flowchart of the circuit of the optimal PI controller.

Fitness Function Definition

The proportion function containing normal effectiveness is the initial step in the process of designing the optimal controller relying on metaheuristic algorithms with favored characteristics and restrictions under the step input signal to reach optimal effectiveness. Some of the major time-related features include overshoot (M_p), rise time (T_r), settling time (T_s), and permanent state error. The minimum value of the proportion function is related to time, and values of the parameters are optimal. Selecting the proportion function is the most important step in applying the algorithm. In general, there are four types of performance: integral absolute error, integral squared error, weighted squared error integral, and weighted absolute error integral at the ITAE performance time. As shown by Jagatheesan and Anand, to confirm the optimal selection of a fitness function for the multi-objective grey wolf optimization algorithm, the ITAE yields the fastest response with minimum overshoot for the methods of optimization, which is the best choice for the optimization of a proportional integral controller [23]. Ansari et al. came to conclusion that utilizing the ITAE as

the fitness function for finding the best gins for traditional PI is in fact to select the multi-objective optimizer. In the study conducted by Kishnani et al., another optimization technique, which is known as the (election-survey optimization), was used. In this technique, the ISE is chosen to obtain the PI gains [33]. Though the gain resulting from merging the errors has been utilized as the assessor of performance, a weighted ISE and an overshoot was considered by Kumar et al., leading to the results in which the overshoot values were above the permitted level. In this paper, the ISE, IAE, and ITSE were applied for confirming the better case. This problem can be mathematically represented as follows [33], [34]:

$$ISE = \int_0^\infty \delta^2(t)dt \tag{28}$$

$$IAE = \int_0^\infty \left|\delta(t)\right| dt \tag{29}$$

$$ITSE = \int_0^\infty t\delta^2(t)dt \tag{30}$$

$$ITAE = \int_0^\infty t\left|\delta(t)\right| dt \tag{31}$$

where τ indicates the upper bound, which is chosen as the steady-state value and δ represents the error of reference to the output. Though the assessment of the old methods is an integrated technique, a novel technique has been suggested recently, in which both the settling time and the overshoot are considered. Therefore, these values must be combined appropriately in the fitness function. This function is defined as follows [33]:

$$Cop = T_s\left(1 + \alpha M_p^2\right) + \beta T_r \tag{32}$$

where MP indicates the overshoot, α represents the adjustable weight for the squared overshoot, and T_S and Tr indicate the settling time and rise time Respectively [21]. To obtain a suitable selection for this weight, the desired percentage must be considered for the overshoot (M%) in which the square root should be unique:

$$\alpha = \left(\frac{1}{M_{\%}}\right)^2 \tag{33}$$

The fitness function, unlike the cost function, is commonly maximized. For this purpose, the following equation is utilized:

$$fit = 1/\mathrm{Cos}\,t \tag{34}$$

Computational Implementation of GWO_AF

Proposed by Mirjalili & Lewis in 2014, the GWO algorithm has been inspired by the grey wolves' leadership and group hunting technique. When designing the GWO algorithm, to transform the hierarchy of the grey wolves into a mathematical model, the readiest solution is considered as the Wolf α. Subsequently, the second and third solution, respectively, are indicated by β and δ. The rest of the suggested solutions are considered as the Wolves ω. The optimization in the GWO algorithm is led by α, β, and δ. Then, these three wolves are followed by the Wolves ω towards the global optimization. In addition to the social leadership, the following equations are used to prompt the surrounding behavior of the grey wolves during the optimization [26].

$$D = \left|C.X_P(t) - X(t)\right| \tag{35}$$

$$X(t+1) = X_P(t) - A.D \tag{36}$$

where t represents the current iteration, C and A are vector coefficients, X_P is the position of the hunting vector, and X indicates the position of the grey wolf's vector. Vectors A and C are calculated as follows:

$$A = 2a.r_1 - 1 \tag{37}$$

$$C = 2.r_2 \tag{38}$$

In this formula, during the periods, the iteration of Component a is linearly reduced from 2 to 0 and r_1 and r_2, which are random vectors within the range of [0, 1] in the GWO algorithm in Fig. 11, find the optimal solutions to the optimization problems by exciting of social leadership and surrounding mechanisms. This algorithm keeps the first three good solutions and forces the other Factors to update their positions, to stimulate, search, and identify the dedicated areas of the search space; the continual optimization process employs the accompanying formulas for any search factor:

$$D_\alpha = \left|C_1.X_a - X\right| \tag{39}$$

$$D_\beta = \left|C_2.X_\beta - X\right| \tag{40}$$

$$D_\delta = \left|C_3.X_\delta - X\right| \tag{41}$$

$$X_1 = X_a - A_1.D_\alpha \tag{42}$$

$$X_2 = X_\beta - A_2.D_\beta \tag{43}$$

$$X_3 = X_\delta - A_3.D_\delta \tag{44}$$

$$X(t+1) = \frac{X_1 - X_2 - X_3}{3} \tag{45}$$

The search by A, the random value of which is above 1 or below −1, has been guaranteed; so that, the search factor is force to deviate from the prey. Another part of the grey wolf optimization algorithm is C, which facilitates the search process. Vector C created a value between 0 and 2 so that [0, 2] is indicative of the prey's random weight for the non-emphasis (C < 1) or statistical emphasis (C > 1) on the prey's impact on the definition of distance. With the help of this, the GWO algorithm will exhibit further random values in the optimization process, resulting in a search that is better than the limited optimum. Here, it is notable that C is not reduced linearly in comparison with A.

To ensure that the search is prioritized not only during the preliminary repetitions but also during the ultimate iterations, it is necessary to use parameter C when generating the random value in every scenario. Particularly in the last iterations, this component is quite useful for the ideal recording.

In the case that |A| < 1, processing with the grey wolf optimization algorithm will start. Whenever Vector A' s random value falls within the [−1, 1] range, the search factor's next position will be somewhere between its current position and the position of the prey. This will facilitate the convergence of the search factor to the estimated position of the prey, which has been given by the problem' alpha, beta, and delta.

The first step in the GWO algorithm is optimizing the random generation of solutions that are part of the preliminary population. Alpha, Beta, and Gamma are the three remaining good solutions that have been maintained as a part of the optimization process. For each Wolf ω (the searcher factor), the position of updating the equations [39] to [45] has been excited; while, Parameters A and a decrease linearly during a period. Therefore, when |A| > 1, the search factors incline toward divergence from hunting but when |A| < 1, they tend toward convergence.

The location and score of Solution Alpha are ultimately returned as the optimum solution found throughout the optimization process after the ultimate criterion is attained. To execute the optimization process of multiple objectives, we combined two new components with it. The MOPSO-like components are those that have been utilized for this particular objective. The initial part of the system is an archive that oversees keeping the optimal solutions that do not include ray domination. The leader's approach, which aids the alpha, beta, and delta in choosing the best option from the archive, makes up the second element [35]. The archive is, in fact, straightforward storage that allows for the safe archiving and retrieval of optimal solutions independent of ray domination. As the best instance of the archives, the controller archive can be mentioned, which can control the archive when a solution is entered or when the archive runs out of capacity. Notice that an archive has several sections. During iteration, the obtained non-dominated solutions are compared with sections of the archive. This might lead to four different cases:

* In the case that the new member has been controlled by at least one of the archive sections, the solution should not be imported into the archive.

* The newly developed solution can regulate any number of previously implemented solutions. In these cases, the controlled solutions in the archive should be eliminated in order that the new solution can be imported into the archive.

* If neither the new solution nor the old one controls the other one, then it is better to add a new solution to the archive.

* If the archive has run out of capacity, the mechanism network should resort the components of the search space, find the most populated piece, and delete its solutions. Subsequently, to advance the ultimate optimal Estimated Pareto front, the novel solution must be imported into the least-populated component.

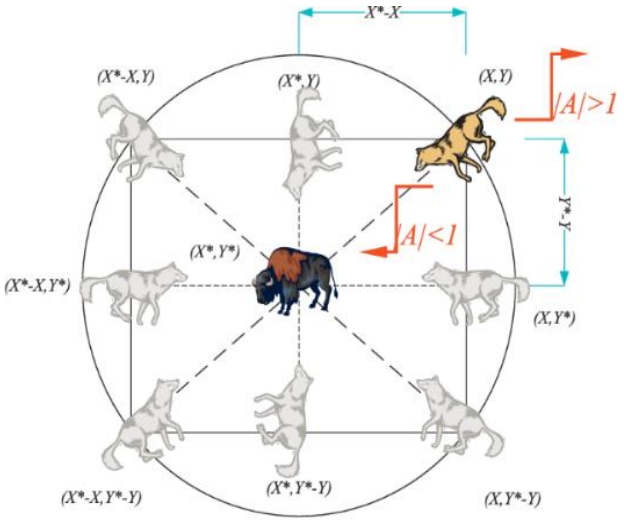


Fig. 11: The search agents' position-updating process and how A affects it [26].

Results and Discussion

The extensive simulation performed in this work aimed to find the interleaved boost converter's dynamic performance. This was done with an optimal proportional-integral controller under SIMULNIK environment in MATLAB. Considering the step input and frequency response speed of the interleaved boost converter with classic controllers, the dynamic performance implies the better dynamic response of the studied converter with GWO_AF-based PI controller in

comparison with the ordinary boost converter.

Table 2: PI parameters

PI parameters	Value
K_p	3^{-3}
K_i	4

The step input response to the interleaved boost converter with optimized PI controller has the fastest response with less overshoot error. It should be noted that the optimization approach is indeed a standard method for designing classic PI controllers; besides, it also exhibits a good performance regarding the close loop of the interleaved boost converter.

The PI controller's parameters, which have been designed using GWO_AF metaheuristic algorithm, are presented in Table 2.

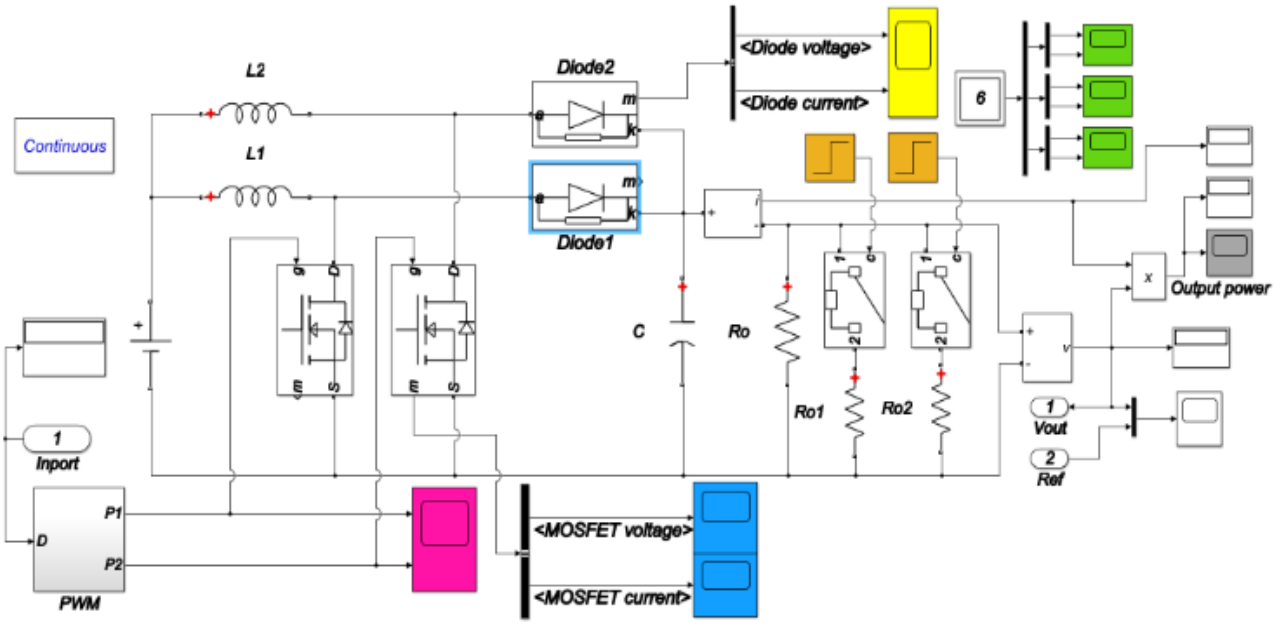


Fig. 12: Schematic of the interleaved boost converter with the proposed controller.

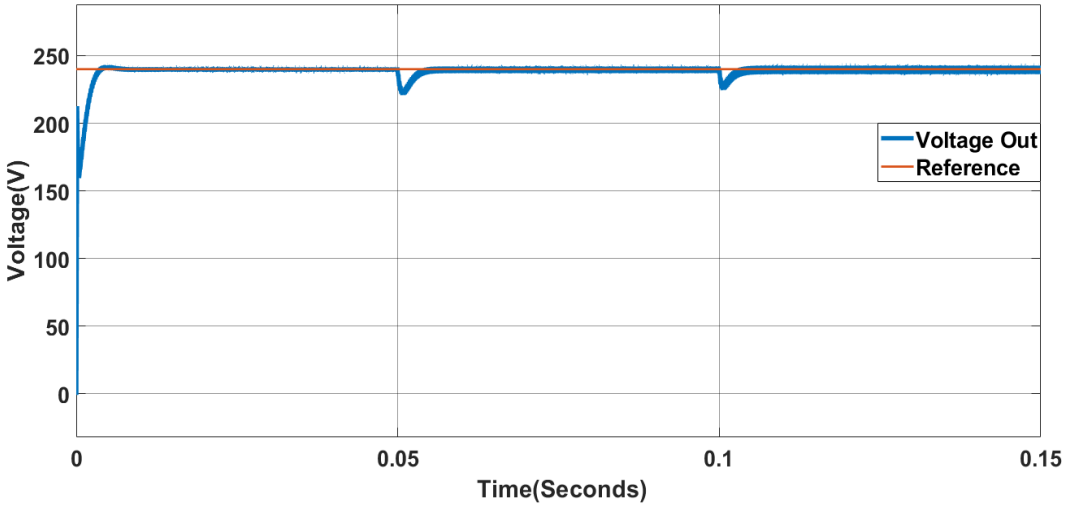


Fig. 13: The waveform of the optimized converter's output voltage.

Fig. 13 shows the optimal output voltage of the converter. After applying a momentary load at times of 0.05 and 0.1 to the converter, the output voltage generally achieves stability compared to the reference voltage 240 by the PI_GWO_AF controller.

Fig. 22 displays the voltage stability and performance against the sudden changes in the load connection to the network when the output voltage increases to 240V for the steady state.

The best closed-loop dynamics (settling time, rise time, and overshoot) exhibited by the GWO_AF-based PI controller with interleaved boost converter were lower than that of the classic PI controller with a boost converter.

In comparison with the boost converter with the optimal controller, the best tracking performance was exhibited by the interleaved boost converter with a PI controller based on the GWO_AF algorithm.

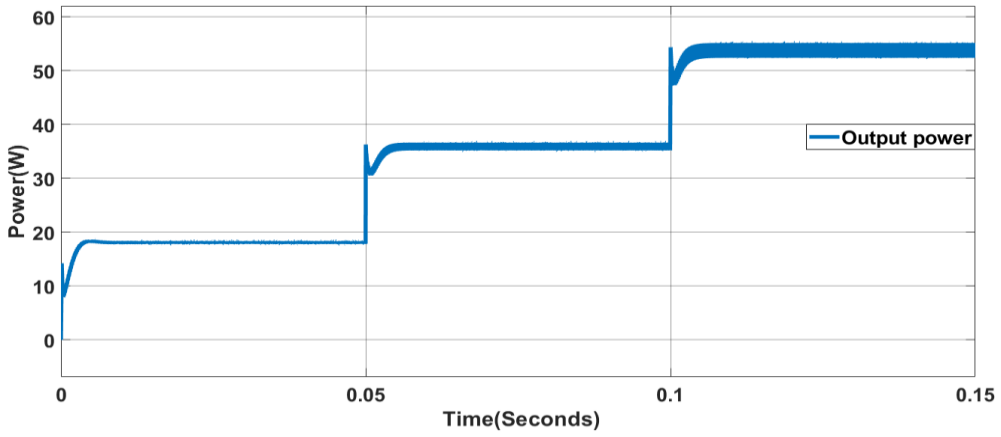


Fig. 14: The waveform of the optimized converter's output power.

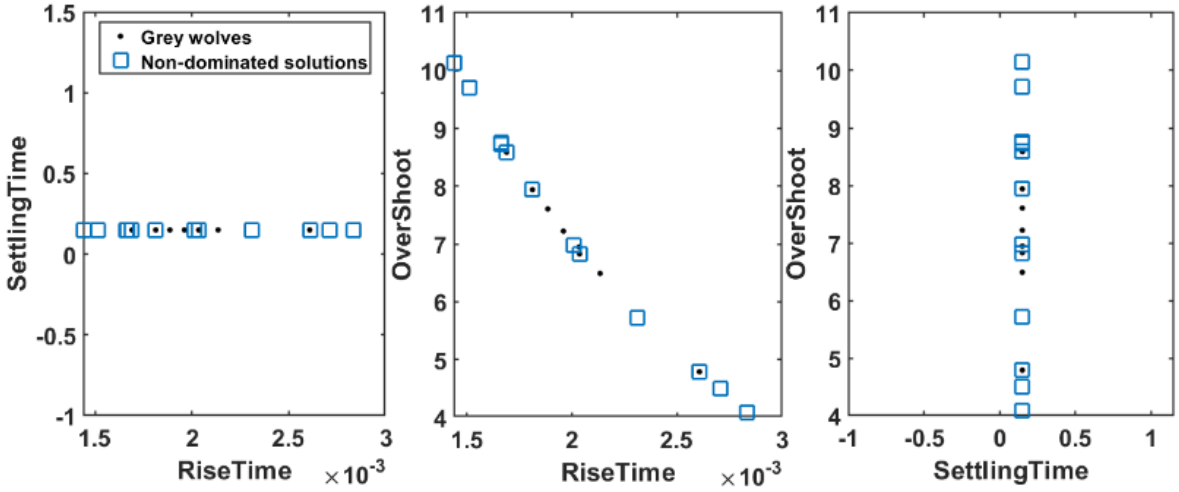


Fig. 15: Estimated Pareto front of the optimized parameters' performance relative to each other.

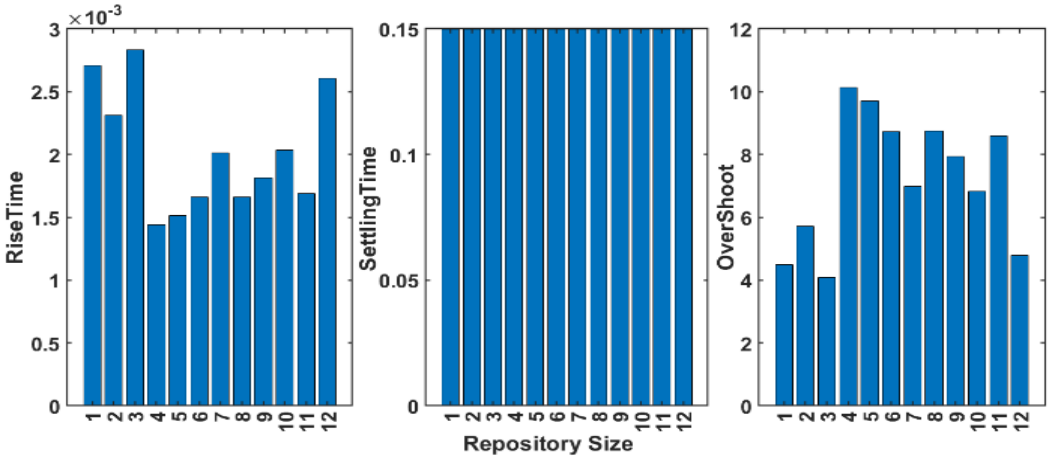


Fig. 16: The bar graph of the optimized parameters' performance in the circuit relative to each other.

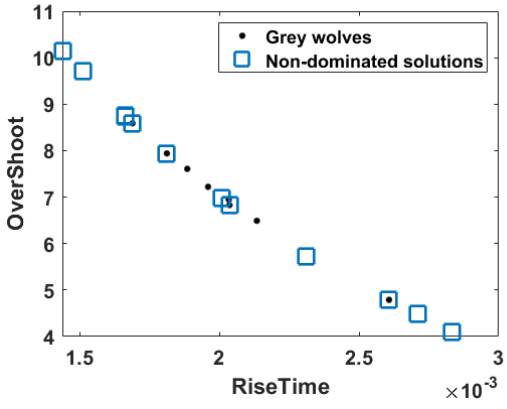


Fig. 17: Estimated Pareto front, Overshoot performance and Rise Time towards each other.

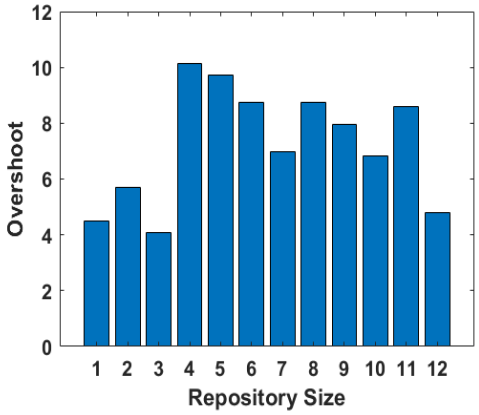


Fig. 19: The optimization performance of the "Overshoot" parameter by the controller.

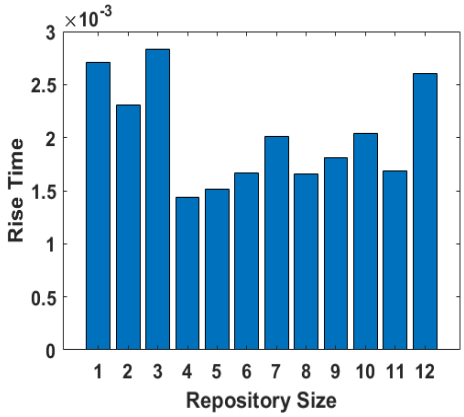


Fig. 18 : The optimization performance of the "Rise Time" parameter by the controller.

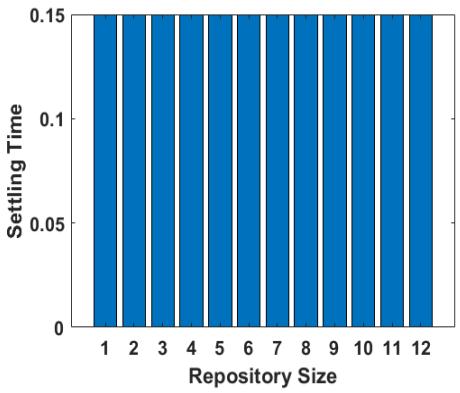


Fig. 20: The optimization performance of the "Settling time" parameter by the controller.

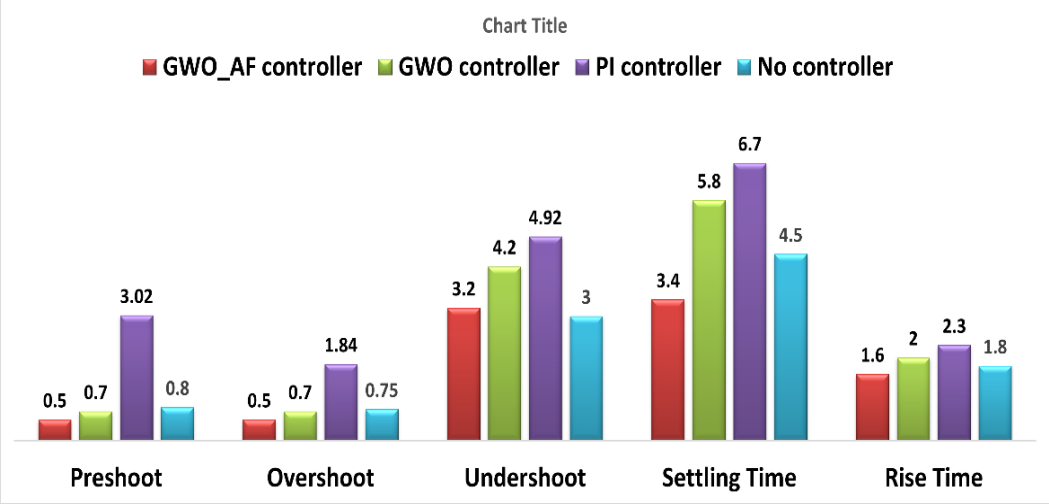


Fig. 21: Comparison of the effect of controller performance on optimized parameter.

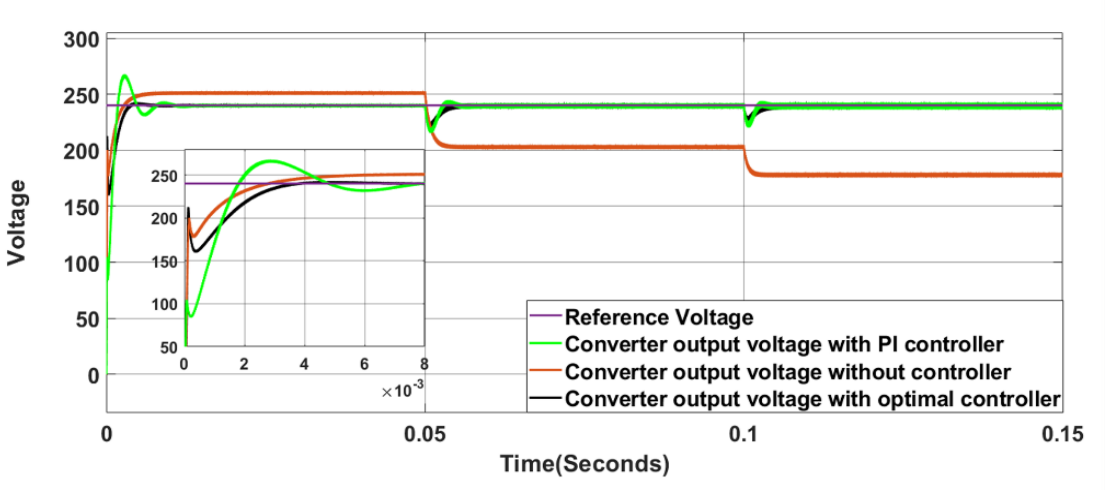


Fig. 22: The Effect of controller performance on the output voltage of converter.

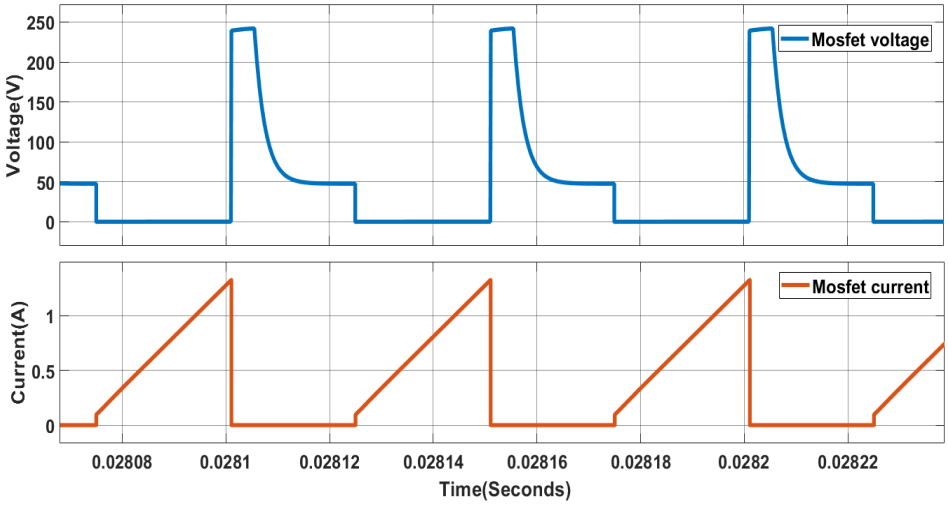


Fig. 23: Current waveform and MOSFET output voltage.

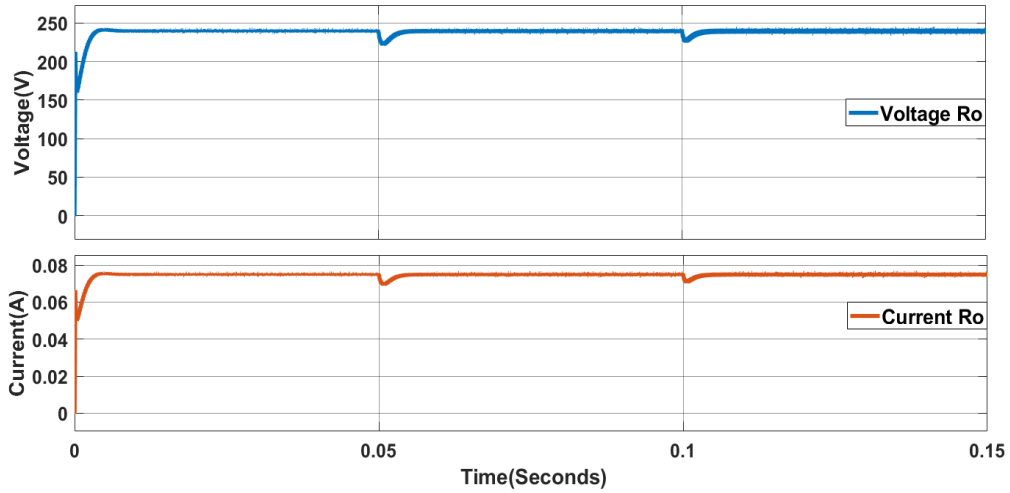


Fig. 24: Current waveform and output resistance voltage.

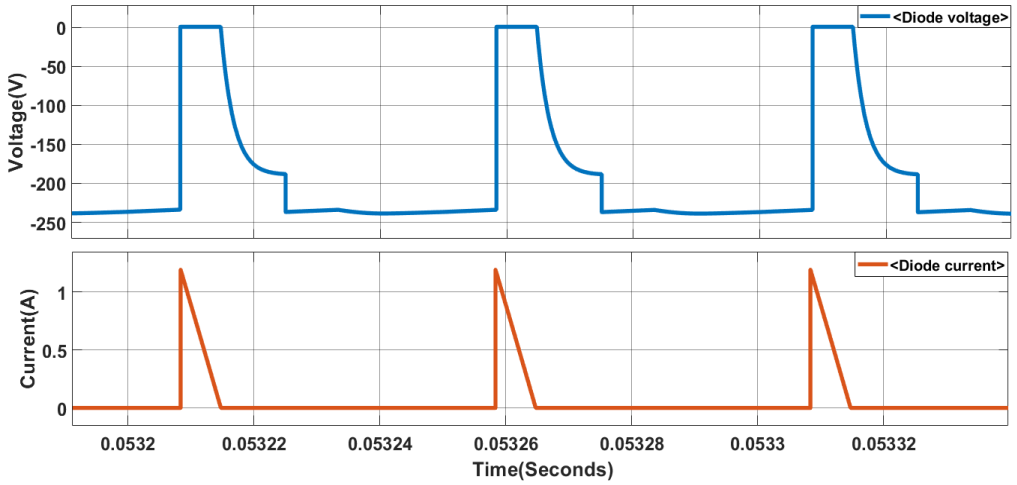


Fig. 25: Current waveform and Diode output voltage.

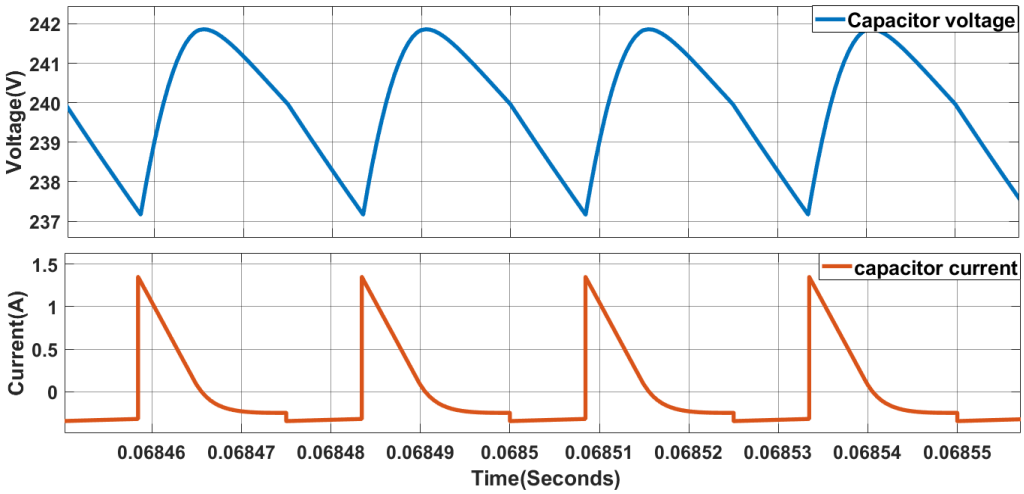


Fig. 26: Output capacitor current waveform.

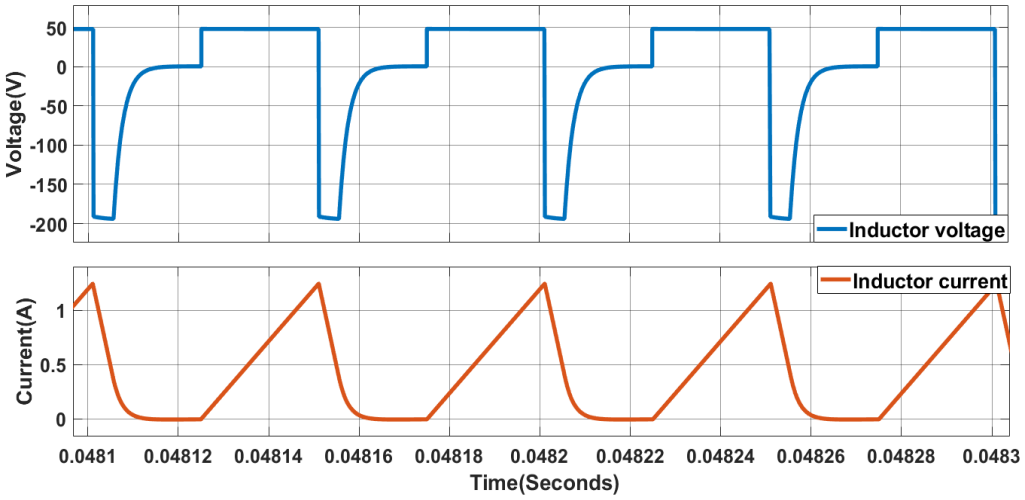


Fig. 27: Current waveform and inductor output voltage.

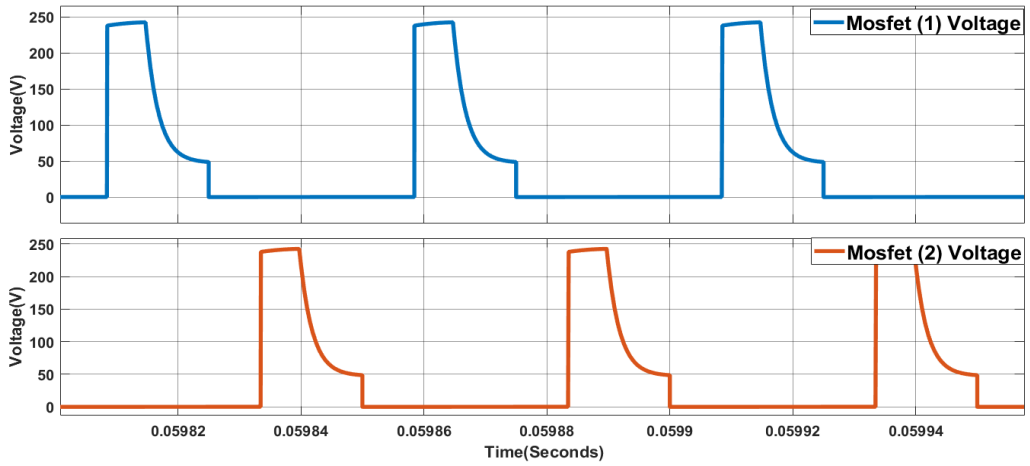


Fig. 28: MOSFET output voltage waveform.

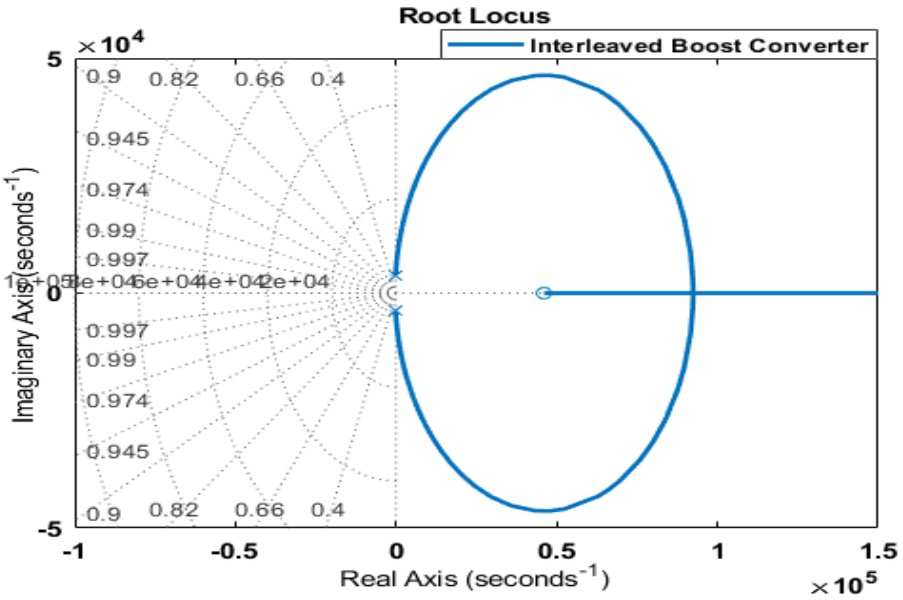


Fig. 29: The pole and zero of the interleaved boost converter.

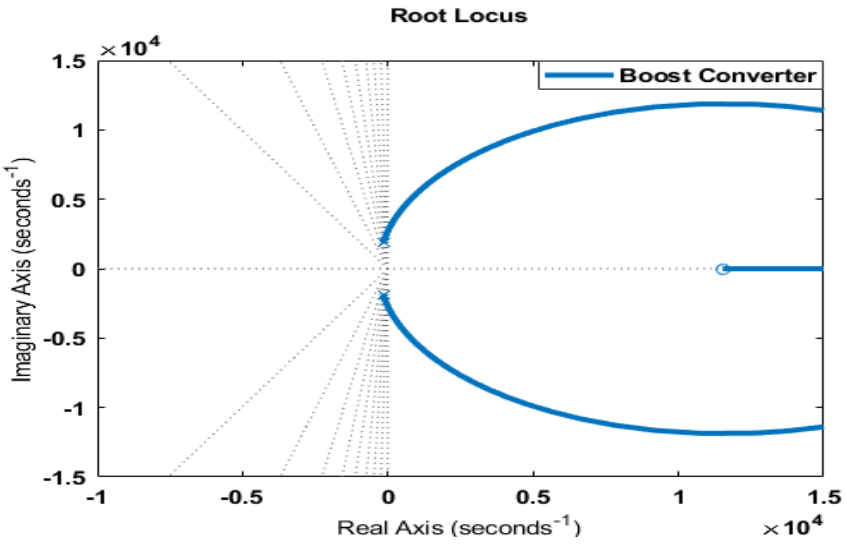


Fig. 30: The pole and zero of the boost converter.

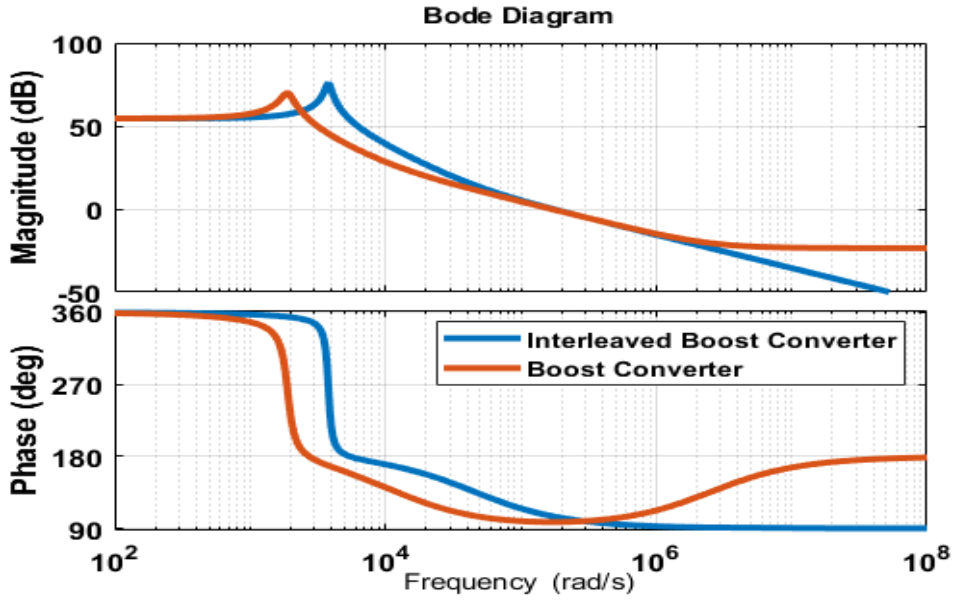


Fig. 31: The bode diagram of the interleaved boost converter.

In the optimal controller based interleaved boost controller, the load voltage followed the reference voltage (Ref), was set rapidly, and preserved the zero steady-state error. Therefore, the simulation results are indicative of the satisfactory tracking performance of the "interleaved boost converter" with optimal controller and its good load adjustment in the closed-loop mode. The simulated response of the input current, the pulse gate for S1, the currents of the Inductors i_{L1} and i_{L2}, and the pulse gate for S2 along the steady state are shown in Fig. 27. As can be inferred from Fig. 28, the inductor's two currents and the two pulse gates are located at a 180◦ angle relative to each other. In comparison with the current of the single inductor, the input current, which results from the inductor's two currents, is made of a smaller ripple. Fig. 29 shows the location of the zero and pole of the interleaved boost converter based on (27). As can be seen, the stability of the proposed converter controller is superior to that of the boost converter (shown in Fig. 30).

Conclusion

This study is aimed at designing and executing a two-phase interleaved boost converter based on an optimal PI controller. This converter is going to be designed and executed for decomposable applications. For this purpose, the controller was initially designed using the classic technique.

Subsequently, the controller's parameters were adjusted by means of the GWO_AF-based optimization method aiming to obtain better performance and stability. Then, simulations were performed to investigate

the interleaved boost converter in terms of comparative analysis and closed-loop performance. Regardless of the operating cycle, the obtained results indicated the lower ripple content of the studied converter in the inductor, input current, and output voltage I comparison with the boost converter.

According to the results, it could be concluded that the proposed converter has exhibited the highest system bandwidth, the best closed-loop performance, and the largest stability margin compared to the converter's open-loop mode. Accordingly, it was decided to utilize the GWO_AF-based PI controller for designing and executing the SMPS by the interleaved boost converter in order that the overall stability and closed-loop performance can be improved. The proposed optimal controller can be utilized for different applications with high degree power converters, including hybrid electrical vehicles and MES, which has not been reported before and has been introduced in this paper for the first time. It can be also executed for the extraction of the maximum power and other objectives. Notably, the proposed converter with very fast transient response and without high-frequency ripples can be also executed for the extraction of the maximum power from the SPV panel.

Author Contributions

S.M Naji Esfahani simulated the converter and controller in MATLAB.

S. H. Zahiri and M. Delshad have supervised the performance optimization and converter controller design. All authors discussed the results and contributed to the final manuscript.

Acknowledgment

We sincerely thank the respected referees for their accurate reviewing of this paper.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

PI	Proportional Integral
CCM	Continuous Conduction Mode
GWO_AF	Grey Wolf Optimization with Aggregation Function
PWM	Pulse Width Modulation

References

[1] R. Giral, L. Martinez-Salamero, S. Singer, "Interleaved converters operation based on cmc," IEEE Trans. Power Electron., 14(4): 643-652, 1999.

[2] P. W. Lee, Y. S. Lee, D. K. Cheng, X. C. Liu, "Steady-state analysis of an interleaved boost converter with coupled inductors," IEEE Trans. Ind. Electron., 47(4): 787-795, 2000.

[3] H. B. Shin, E. S. Jang, J. G. Park, H. W. Lee, T. Lipo, "Small-signal analysis of multiphase interleaved boost converter with coupled inductors," IEE Proc.: Electr. Power Appl., 152(5): 1161-1170, 2005.

[4] C. Wang, "Investigation on interleaved boost converters and applications," Ph.D. thesis, Virginia Tech, 2009.

[5] Z. C. Zhang, B. T. Ooi, "Multimodular current-source SPWM converters for superconducting magnetic energy storage system: Sinusoidal pulse-width modulation," IEEE Trans. Power Electron., 8(3): 250-256, 1993.

[6] Y. Sumi, Y. Harumoto, T. Hasegawa, M. Yano, K. Ikeda, T. Matsuura, "New static VAR control using force-commutated inverters," IEEE Trans. Power Appar. Syst., 9: 4216-4224, 1981.

[7] Z. Zhang, J. Kuang, X. Wang, B. T. Ooi, "Force commutated HVDC and SVC based on phase-shifted multi-converter modules," IEEE Trans. Power Delivery, 8(2): 712-718, 1993.

[8] K. K. Hedel, "High-density avionic power supply," IEEE Trans. Aerosp. Electron. Syst., 5: 615-619, 1980.

[9] T. Mishima, Y. Takeuchi, M. Nakaoka, "Analysis, design, and performance evaluations of an edge-resonant switched capacitor cell-assisted soft-switching PWM boost DC-DC converter and its interleaved topology," IEEE Trans. Power Electron., 28(7): 3363-3378, 2012.

[10] P. V. Nandankar, J. P. Rothe, "Highly efficient discontinuous mode interleaved DC-DC converter," in Proc. 2016 IEEE International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT): 476-481, 2016.

[11] F. Yang, X. Ruan, Y. Yang, Z. Ye, "Interleaved critical current mode boost PFC converter with coupled inductor," IEEE Trans. Power Electron., 26(9): 2404-2413, 2011.

[12] M. Pavlovsky, G. Guidi, A. Kawamura, "Assessment of coupled and independent phase designs of interleaved multiphase buck/boost DC-DC Converter for EV power train," IEEE Trans. Power Electron., 29(6): 2693-2704, 2013.

[13] Y. S. Roh, Y. J. Moon, J. Park, C. Yoo, "A two-phase interleaved power factor correction boost converter with a variation-tolerant phase shifting technique," IEEE Trans. Power Electron., 29(2): 1032-1040, 2013.

[14] S. Talebi, E. Adib, M. Delshad, "A High-Gain interleaved DC-DC converter with passive clamp circuit and low current ripple," Iran. J. Sci. Technol. Trans. Electr. Eng., 45(1): 141-153, 2021.

[15] B. C. Barry, J. G. Hayes, M. S. Rylko, "CCM and DCM operation of the interleaved two-phase boost converter with discrete and coupled inductors," IEEE Trans. Power Electron., 30(12): 6551-6567, 2014.

[16] J. K. Seok, A. Parastar, "Modeling and control of the average input current for three-phase interleaved boost converters," IEEE Trans. Ind. Appl., 51(3): 2340-2351, 2014.

[17] S. Banerjee, A. Ghosh, N. Rana, "Design and fabrication of closed loop two-phase interleaved boost converter with type-III controller," in Proc. the 42nd Annual Conference of the IEEE Industrial Electronics Society (IECON): 3331-3336, 2016.

[18] M. A. Johnson, M. H. Moradi, PID control, Springer, 2005.

[19] S. Z. Zhao, M. W. Iruthayarajan, S. Baskar, P. N. Suganthan, "Multiobjective robust PID controller tuning using two lbests multi-objective particle swarm optimization," Inf. Sci., 181(16): 3323-3335, 2011.

[20] N. S. Shahraki, S. H. Zahiri, "Multi-objective optimization algorithms in analog active filter design," in Proc. 2020 8th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS): 105-109, 2020.

[21] H. Freire, P. Moura Oliveira, E. Solteiro Pires, "From single to many-objective PID controller design using particle swarm optimization," Int. J. Control Autom. Syst., 15(2): 918-932, 2017.

[22] A. Ghosh, S. Banerjee, M. K. Sarkar, P. Dutta, "Design and implementation of type-ii and type-iii controller for dc-dc switched-mode boost converter by using k-factor approach and optimisation techniques," IET Power Electr., 9(5): 938-950, 2016.

[23] R. Eberhart, J. Kennedy, "A new optimizer using particle swarm theory," in Proc. the Sixth International Symposium on Micro Machine and Human Science (MHS'95): 39-43, 1995.

[24] F. I. Chou, Y. C. Cheng, P. Y. Yang, J. T. Tsai, J. H. Chou, "Optimal multiobjective PID design by PSO," in Proc. 2019 IEEE International Conference on Systems, Man, and Cybernetics (SMC): 927-928, 2019.

[25] L. C. Borin, E. Mattos, C. R. Osorio, G. G. Koch, V. F. Montagner, "Robust PID controllers optimized by PSO algorithm for power converters," in Proc. 2019 IEEE 15th Brazilian Power Electronics Conference and 5th IEEE Southern Power Electronics Conference (COBEP/SPEC): 1-6, 2019.

[26] S. Mirjalili, J. S. Dong, "Multi-objective grey wolf optimizer," in Multi-Objective Optimization using Artificial Intelligence Techniques, Springer, pp. 47-58, 2020.

[27] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, H. Alhussian, M. G. Ragab, A. Alqushaibi, "Binary multi-objective grey wolf optimizer for feature selection in classification," IEEE Access, 8: 106247-106263, 2020.

[28] S. Banerjee, A. Ghosh, N. Rana, "An improved interleaved boost converter with PSO-based optimal type-III controller," IEEE J. Emerging Sel. Top. Power Electron., 5(1): 323-337, 2016.

[29] S. Mukherjee, "A SEPIC-Cuk-CSCCC based SIMO converter design using PSO-MPPT for renewable energy application," J. Electr. Comput. Eng. Innov., 10(2): 437-446, 2022.

[30] K. L. Shenoy, C. G. Nayak, R. P. Mandi, et al., "Design and implementation of interleaved boost converter," Int. J. Eng. Technol. (IJET), 9(3S): 496-502, 2017.

[31] K. Eltag, M. S. Aslamx, R. Ullah, "Dynamic stability enhancement using fuzzy PID control technology for power system," Int. J. Control Autom. Syst., 17(1): 234-242, 2019.

[32] A. Husna, M. Mat, M. Yasin, M. Jusoh, Y. Irwan, M. A. Rahim, S. M. Esa, "Critical review: Adaptive pole assignment PID controller on DC-DC converters," in Proc. IOP Conference Series: Materials Science and Engineering, 767: 012039, 2020.

[33] K. L. Shenoy, C. G. Nayak, R. P. Mandi, et al., "Design and implementation of interleaved boost converter," Int. J. Eng. Technol. (IJET), 9(3S): 496-502, 2017.

[34] E. D. P. Puchta, H. V. Siqueira, M. dos Santos Kaster, "Optimization tools based on metaheuristics for performance enhancement in a Gaussian adaptive PID controller," IEEE Trans. Cybern., 50(3): 1185-1194, 2019.

[35] I. Behravan, S. H. Zahiri, S. M. Razavi, R. Trasarti, "Clustering a big mobility dataset using an automatic swarm intelligence-based clustering method," J. Electr. Comput. Eng. Innov., 6(2): 251-271, 2018.

[36] G. H. Valencia-Rivera, I. Amaya, J. M. Cruz-Duarte, J. C. Ortíz-Bayliss, J. G. Avina-Cervantes, "Hybrid controller based on LQR applied to interleaved boost converter and microgrids under power quality events," Energies, 14(21): 6909, 2021.

[37] A. Tjahjono, D. O. Anggriawan, M. N. Habibi, E. Prasetyono, "Modified grey wolf optimization for maximum power point tracking in photovoltaic system under partial shading conditions," Int. J. Electr. Eng. Informatics, 12(1): 94-104, 2020.

[38] C. Komathi, M. G. Umamaheswari, "Design of gray wolf optimizer algorithm-based fractional order PI controller for power factor correction in SMPS applications," IEEE Trans. Power Electron., 35(2): 2100-2118, 2019.

[39] H. M. H. Farh, A. M. Eltamaly, M. S. Al-Saud, "Interleaved boost converter for global maximum power extraction from the photovoltaic system under partial shading," IET Renew. Power Gener., 13(8): 1232-1238, 2019.

[40] J. Y. Shi et al., "Dual-algorithm maximum power point tracking control method for photovoltaic systems based on grey wolf optimization and golden-section optimization," J. Power Electron., 18(3): 841-852, 2018.

[41] J. Jayaudhaya, K. Ramash Kumar, V. Tamil Selvi, N. Padmavathi, "Improved performance analysis of PV array model using flower pollination algorithm and gray wolf optimization algorithm," Math. Probl. Eng., 5803771: 1-15, 2022.

Biographies

Seyed Mohammad Naji Esfahani received the B.S. and M.S. degrees in electrical engineering in 2015 and 2017 from University of Applied Science Technology Jahad Daneshgahi Isfahan and Islamic Azad University of Isfahan (Khorasgan) Branch respectively. He is currently pursuing the Ph.D. degree in the Department of Electronics Engineering, University of Birjand, Iran. His research interest includes soft switching techniques in DC-DC converters and optimization of control systems by meta-heuristic algorithms.

- Email: sm.naji@birjand.ac.ir
- ORCID: 0000-0003-1925-533X
- Web of Science Researcher ID: NA
- Scopus Author ID:NA
- Homepage: NA

Seyed Hamid Zahiri received the B.Sc., M.Sc., and Ph.D. degrees in Electronics Engineering respectively from Sharif University of Technology, Tehran, Tarbiat Modarres University, Tehran, and Ferdowsi University, Mashhad, Iran, in 1993, 1995, and 2005. Currently, he is Professor with the Department of Electronics Engineering, University of Birjand, Birjand, Iran. His research interests include pattern recognition, evolutionary algorithms, swarm intelligence algorithms, and soft computing.

- Email: hzahiri@birjand.ac.ir
- ORCID: 0000-0002-1280-8133
- Web of Science Researcher ID: NA
- Scopus Author ID:NA
- Homepage: NA

Majid Delshad was born in Isfahan, Iran, in 1979. He received the B.S. degree in electrical engineering from Kashan University, in 2001, and the M.S. degree in electrical engineering, in 2004, and the Ph.D. degree in electrical engineering from the Isfahan University of Technology, Iran, in 2010. He is an Associate Professor with Isfahan (Khorasgan) Branch, IAU. His research interests include soft-switching techniques in DC_DC converters and current-fed converters.

- Email: delshad@khuisf.ac.ir
- ORCID: 0000-0002-2637-5965
- Web of Science Researcher ID: NA
- Scopus Author ID: 24721046500
- Homepage: NA

How to cite this paper:
S. M. Naji Esfahani, S. H. Zahiri, M. Delshad, "Application of grey wolf optimization algorithm with aggregation function on designing interleaved boost converter," J. Electr. Comput. Eng. Innovations, 12(1): 39-56, 2024.

DOI: 10.22061/jecei.2023.9355.610

URL: https://jecei.sru.ac.ir/article_1895.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Predicting the Sentiment of Tweet Replies Using Attentive Graph Convolutional Neural Networks

S. Nemati *

Department of Computer Engineering, Shahrekord University, Shahrekord, Iran.

Article Info

Article History:
Received 18 April 2023
Reviewed 20 May 2023
Revised 09 July 2023
Accepted 15 August 2023

Keywords:
Sentiment analysis
Deep leaning
Social media
Twitter
Graph convolutional neural networks

*Corresponding Author's Email Address: *s.nemati@sku.ac.ir*

Abstract

Background and Objectives: Twitter is a microblogging platform for expressing assessments, opinions, and sentiments on different topics and events. While there have been several studies around sentiment analysis of tweets and their popularity in the form of the number of retweets, predicting the sentiment of first-order replies remained a neglected challenge. Predicting the sentiment of tweet replies is helpful for both users and enterprises. In this study, we define a novel problem; given just a tweet's text, the goal is to predict the overall sentiment polarity of its upcoming replies.

Methods: To address this problem, we proposed a graph convolutional neural network model that exploits the text's dependencies. The proposed model contains two parallel branches. The first branch extracts the contextual representation of the input tweets. The second branch extracts the structural and semantic information from tweets. Specifically, a Bi-LSTM network and a self-attention layer are used in the first layer for extracting syntactical relations, and an affective knowledge-enhanced dependency tree is used in the second branch for extracting semantic relations. Moreover, a graph convolutional network is used on the top of these branches to learn the joint feature representation. Finally, a retrieval-based attention mechanism is used on the output of the graph convolutional network for learning essential features from the final affective picture of tweets.

Results: In the experiments, we only used the original tweets of the RETWEET dataset for training the models and ignored the replies of the tweets in the training process. The results on three versions of the RETWEET dataset showed that the proposed model outperforms the LSTM-based models and similar state-of-the-art graph convolutional network models.

Conclusion: The proposed model showed promising results in confirming that by using only the content of a tweet, we can predict the overall sentiment of its replies. Moreover, the results showed that the proposed model achieves similar or comparable results with simpler deep models when trained on a public tweet dataset such as ACL 2014 dataset while outperforming both simple deep models and state-of-the-art graph convolutional deep models when trained on the RETWEET dataset. This shows the proposed model's effectiveness in extracting structural and semantic relations in the tweets.

Introduction

Twitter is a social networking and microblogging platform with nearly 400 million users worldwide. Its interesting characteristics have made Twitter a prominent communication tool for not only ordinary people but also particular users, including students [1], [2], politicians [3], [4] medical specialists [5], [6], athletes [7] and traders [8],

[9]. Users post their tweets to share daily updates, talk about their opinion and emotions, and keep in contact with friends and family [10], [11]. Moreover, people use Twitter to connect with others to discuss common interests and concerns with other users worldwide. This may be achieved by tweeting, retweeting, mentioning, or replying to other users' tweets [11]. Therefore, in recent years, mining Twitter for information about people and their opinion, sentiment, preferences, and reactions to events has attracted the increasing attention of researchers, companies, and media organizations [10].

Many research studies have addressed Twitter sentiment analysis since 2011 [10]. The main goal of Twitter sentiment analysis is to detect the sentiment polarity of a given tweet in terms of positive, negative, or neutral [11], [12]. In addition to standard tweet polarity detection, more fine-grained tasks, including emotion detection [13], personality detection [14], event detection [15], stock prediction [16], and election prediction [17], have also been investigated in recent years. Sentiment analysis of Twitter data has more challenges than sentiment analysis of reviews or similar texts [12], [18]. This has several reasons, including the limited length of tweets, the use of informal language and unique abbreviations, and complex relations formed using the mention, retweet, and reply mechanisms of Twitter. Therefore, several studies addressed sentiment analysis of Twitter data for English and other languages in the last decade [19]-[22].

With the growing number of Twitter users and the increase of tweets' impact, users' desire to capture other users' attention via their tweets has increased [23]. High-quality tweets (i.e., those that capture others' attention) can increase users' reputations [24]. Therefore, predicting other users' reactions to a tweet is essential for users, especially before they post their tweets [24], [25]. The number of times others like a tweet or, similarly, the number of retweets may be signs of a good impression and hence may be used as popularity metrics for a tweet [24]. In addition to these metrics, tweet replies may also be analyzed to detect the sentiment of repliers expressed in their replies as a measure of popularity for the source tweet. To predict other users' reactions to a tweet before posting, it is necessary to analyze the textual content of the tweet. This is a challenging problem among natural language processing tasks because tweets have limited length, forcing users to abbreviate words, invent acronyms on the fly, or even omit words [10], [26].

Every tweet may produce positive, negative, or neutral sentiments and reactions in its readers [10]. Such responses may be shown in terms of likes or dislikes, retweets, or posting textual replies. For likes and retweets, the number of users who like a tweet or retweet it may be considered a factor for measuring the positive reaction of other users [24]. However, for replies, the number of replies does not necessarily show the popularity and the positive responses of others. In this case, the tweet replies' content must be considered to determine how positive/negative the reactions are [27]. Several studies have addressed the problem of predicting the number of likes and retweets in recent years [23], [28]-[30]. These studies usually model the issue as a regression problem in which the model's output is the predicted number of likes or retweets over time [29]. However, indicating other users' sentiments shown in their replies has been neglected in previous studies [27].

Recently, Arasteh et al. [27] addressed the problem of predicting the overall sentiment of tweet replies and proposed a deep learning-based method for this problem. Specifically, they created a relatively large dataset of tweets and their first-order replies, RETWEET, and trained a bi-directional long short-term memory (Bi-LSTM) deep model on manually labeled tweets from the SemEval datasets [31]. Then, using this trained model, they predict the sentiment polarity of all tweet replies without considering the source tweets. Finally, using a heuristic averaging algorithm, they assigned a label to each source tweet according to its replies' polarity labels [27]. Although this study presented the problem of predicting tweet replies' sentiment for the first time, the main shortcoming is the need for having all replies for labeling a tweet. Ignoring the content of the source tweet and assigning a sentiment polarity label using its replies necessitate waiting for others' reactions in terms of their reply to predict the overall sentiment of replies. This seems to be the main weakness of their proposed solution to the problem [27].

In this study, we define a new problem as follows. Given only the textual content of a source tweet, the task is to predict the overall sentiment polarity of upcoming replies. To address this problem, we propose a new deep learning-based model for processing tweets' textual content and predicting their replies' overall sentiment. To this end, we used the RETWEET dataset [32], which contains several tweets and their corresponding first-order replies. In our proposed model, unlike [32], we do not use replies' textual content and only exploit the source tweets' content. Specifically, we trained a graph convolutional network (GCN) on the textual content of source tweets to learn the structural and semantic relations in the text. Then, we evaluate the trained network on unseen tweets in the dataset. In summary, the main contributions of the current study are as follows:

- Defining the problem of predicting the overall sentiment polarity of tweet replies only based on the textual content of the source tweet.
- Proposing a graph convolutional network model for exploiting structural and semantic relations in the tweets.

• Comparing the baseline and state-of-the-art graph convolutional network models with the proposed model on three versions of the RETWEET dataset.

The remainder of the paper continues as follows. In the next section, a brief overview of related studies will be presented. The proposed model will be described in section III. Experimental results are shown in section IV. Conclusions and directions for future work will be discussed in the last section.

Literature Review

In this section, a brief overview of related studies is presented in two subsections as follows. Some Twitter data analysis studies are shown in the first subsection, and deep learning-based models for sentiment analysis are presented in the following subsection.

A. Twitter Data Analysis

Kouloumpis et al. [10] investigated using linguistic features for message-level tweet sentiment analysis. They used a machine learning method and utilized lexical resources and hashtag information in training. They showed that part-of-speech (POS) features were not helpful, while sentiment linguistic features and emoticons are helpful for classification [10]. Agarwal et al. [33] proposed a machine learning approach for sentiment analysis of Twitter data and modeled the problem as binary and 3-way classification problems. They evaluated unigram, feature-based, and tree-based models and showed that the combination of these models outperformed the baseline and each model in isolation [33]. Mohammad et al. [34] designed two sentiment lexicons and proposed a machine learning-based classifier for message-level and term-level sentiment classification of Twitter data. They showed that their lexicon-based approach outperformed the machine learning-based method.

Some recent studies investigated problems that use sentiment analysis to address other issues. For example, Abdar et al. [26] proposed a model for detecting people's attitudes toward energy in Alaska. They used Twitter as a data source in which people express their sentiments and emotion towards different subjects, including Energy. Gagne et al. [1] analyzed nursing student tweets in three countries during COVID-19. They investigated the opinion of students in their tweets to help nurse educators better understand the students. Basiri et al. [12] proposed a deep learning-based model for sentiment analysis of tweets in eight countries during the COVID-19 pandemic. They offered a fusion model and showed that the sentiment intensity expressed by people at different times and governments was not identical. Ali et al. [35] proposed a deep learning model for sentiment analysis of tweets in Pakistan. They aimed to predict the results of the Pakistan general election in 2018 using Twitter data.

Hong et al. [36] investigated the problem of predicting the popularity of tweets and used the number of retweets as the measure of popularity. They employed tweet contents and metadata of tweets, including temporal data and user data. In a similar study, Petrovic et al. [29] investigated the problem of predicting the number of retweets and proposed a machine, learning-based model. They showed that although social features performed very well, tweet features could also be used in the model to reach human-level accuracy. Daga et al. [24] evaluated some machine learning methods learned on a bag-of-words model and word embedding features to predict the number of likes and retweets for a source tweet. They showed that bag-of-words features were more helpful than embedding features for this task [24].

Lou et al. [37] introduced the problem of predicting the users who retweet a source tweet. They proposed a machine learning-based method and used features such as retweet history and followers-related features. They showed that common interests and the history of retweeting were factors that could be used for predicting future retweets [37]. Wang et al. [38] proposed a deep learning-based model to analyze users' retweeting behavior. They integrated user-based and message-based features to model the group retweeting behavior and tweets' content. In a similar study, Firdaus et al. [39] explored the problem of the retweeting behavior of users and focused on the topic's impact. Specifically, they investigated the effect of a user's topic-related sentiment on their retweet decision. They concluded that the topic and users' sentiment toward the topic were important for modeling their retweet behavior [39].

Some recent studies addressed the problem of tweet popularity prediction using novel approaches. For example, Lymperopoulos [40] proposed a model based on electronic circuits for predicting the popularity of tweets in terms of their number of retweets. As another example, Garvey et al. [41] proposed an artificial intelligence probabilistic model for generating popular tweets. Specifically, they used econometrics, machine learning, and Bayesian theory to create the structure of high-impact tweets. Gao et al. [28] proposed a heterogeneous bass model for the prediction of the popularity of tweets. They considered tweets with a similar topic, using a clustering approach and linear regression to improve the system's performance. Rivadeneira et al. [23] proposed an evidential reasoning model for predicting tweets' impact. Specifically, they used five features of tweets to indicate the number of electoral-related retweets.

B. Deep Learning

Several studies applied deep learning techniques to sentiment analysis problems in different domains in recent years. For example, as one of the first applications

of the deep model in the sentiment analysis domain, Poria et al. [42] proposed a feature extraction method based on convolutional networks. They used the extracted features for multimodal sentiment analysis of short video clips. They reported a 14% improvement over existing methods for the same task. Edara et al. [43] applied LSTM to the problem of sentiment analysis of cancer-affected patients' tweets. They showed that their deep model outperforms traditional machine learning models. Basiri et al. [44] proposed a 3-way fusion model of deep and conventional learning techniques for sentiment analysis of drug reviews. They showed that their model outperformed traditional and deep models and considered classifier confidence in its decisions. Muhammad Shah et al. [45] proposed a deep model for multimodal patient review sentiment analysis. They processed both textual and image content of patients' reviews published on the Yelp.com platform.

Parimala et al. [46] proposed an LSTM deep model for sentiment analysis of tweets collected before, after, and during disasters. They compared their model with traditional learning models and reported a slightly better performance for the binary classification scenario. Basiri et al. [12] proposed a deep fusion model consisting of four deep and one traditional learning method for analyzing COVID-19 tweets in different countries. Serrano-Guerrero et al. [47] addressed the problem of sentiment analysis and emotion recognition of patients' reviews. They proposed a hybrid of bidirectional gated recurrent unit (Bi-GRU) and convolutional network to classify reviews. They also evaluated different word embeddings for their models and showed that their clinical-domain word embedding model outperformed other deep and traditional learning models. Basiri et al. [48] proposed a Bi-LSTM model for sentiment analysis of online doctor reviews. They introduced the PODOR dataset containing Persian online doctor reviews and showed that their proposed deep model outperforms traditional learning models for the polarity detection of online doctor reviews.

Some recent studies applied deep learning models to the problem of sentiment analysis in other languages. For example, Shehu et al. [49] evaluated different data augmentation and deep learning models on Turkish tweets. They compared their models with traditional machine learning models and concluded that conventional models outperformed deep models in speed, but deep models performed better. Several studies applied deep learning models to Arabic sentiment analysis [50]. For example, Elfaik et al. [51] used Bi-LSTM, Saleh et al. [52] used a hybrid of CNN and LSTM models, and Al-Dabet et al. proposed a CNN-based model for aspect-based sentiment analysis of Arabic texts. Dashtipour et al. [53] used LSTM and CNN for Persian sentiment analysis of movie reviews. Bokaee Nezhad et al. [54] applied a combination of CNN and LSTM models to COVID-19 tweets in the Persian language. Gonzalez et al. used pre-trained Bert models for Spanish tweet sentiment analysis. Gan et al. used an attention mechanism on a CNN-BiLSTM model for Chinese sentiment analysis.

Smetanin et al. [55] applied a transformer-based deep model to Russian sentiment analysis. In recent years, other languages have also been the target of deep learning methods for sentiment analysis problems. In summary, compared to the previous studies, the novelty of the current research is two-fold. First, we introduce a new problem in the domain of sentiment analysis of Twitter data. Second, we propose a new attentive graph convolutional deep model for solving the problem. The proposed model will be described in the next section in more detail.

Proposed Model

We exploit graph-based convolutional neural networks in the proposed model to consider tweets' structural and semantic information. The overall structure of the proposed model is shown in Fig. 1.

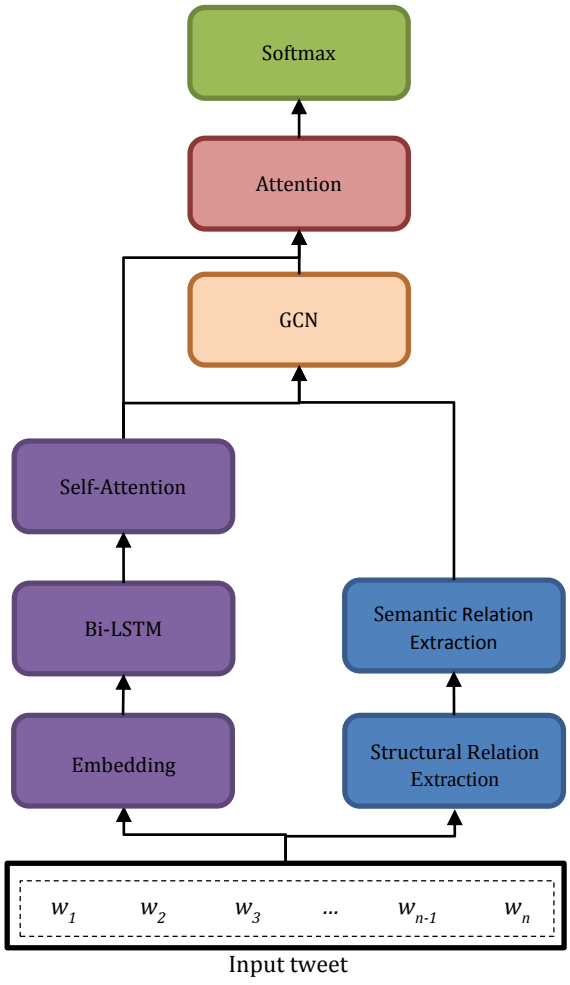


Fig. 1: The overall structure of the proposed model.

There are two parallel branches in the proposed model as follows. The first branch, which starts with the embedding module, extracts the contextual representation of the input tweets. The second branch extracts the structural and semantic information from tweets. Specifically, the input tweets $t = \{w_1, w_2, \dots, w_n\}$, containing *n* words, is sent to both embedding and structural relation extraction modules for converting to a numerical vector and extracting the structural graph, respectively.

A. Contextual Representation Branch

In the embedding module, each tweet is converted to a numerical matrix using a lookup table which is usually derived from transformer-based pre-trained word embeddings such as BERT [56], Elmo [57], or Glove [58]. The numerical representation of each tweet, *t*, contains *n* vectors of length *m*, where *m* is the dimension of the word vectors in the lookup table. In the current study, we used 300-dimensional vectors of Glove trained on 42 billion words from Wikipedia pages and newswires as the lookup table [58].

The Bi-LSTM module takes the embedding matrix of each tweet as input and derives its hidden contextual representations as follows.

$$H^c = \{h_1^c, h_2^c, \dots, h_n^c\} = Bi - LSTM(x) \tag{1}$$

where,

$$x = [x_1, x_2, \dots, x_n] \text{ and } x_i \in \mathbb{R}^m \tag{2}$$

The self-attention module is used on top of the Bi-LSTM module to learn syntactical dependencies [59], [60]. Using this module, each word in the tweet pays attention to other words regardless of their position. To achieve this, three parameters, namely *Q* (queries), *K* (keys), and *V* (values), are combined as follows [59]:

$$Att(Q, K, V) = softmax(QK^T)V \tag{3}$$

To obtain the values of the above three parameters, three randomly initialized weight matrices, W_Q, W_K, W_V and the input of the self-attention module, which is here the output of the Bi-LSTM module, are used as follows:

$$\begin{aligned} SelfAtt(H^c) &= Att(H^c\mathrm{W}_Q, H^c\mathrm{W}_K, H^c\mathrm{W}_V)V \\ &= softmax(H^c\mathrm{W}_{QK}H^{c^T})H^cW_V \end{aligned} \tag{4}$$

where $\mathrm{W}_{QK} = \mathrm{W}_Q W_K^T$.

B. Relation Extraction Branch

The structural relation extraction module is used in the proposed method to construct the dependency graph of tweets. To this aim, we first build the dependency tree of an input tweet using the SpaCy module [61]. Then, we make the adjacency matrix $D \in \mathbb{R}^{n \times n}$ of the tweet using the dependency tree by setting $D_{i,j}$ to one if there is a dependency between the i^{th} and j^{th} words and assigning it to zero otherwise. This strategy is proposed in [62] to create an undirected dependency graph. An illustrative example of converting a sample tweet "*Some universities charge huge fees*" to its corresponding adjacency matrix, is shown in Fig. 2. As shown in the figure, the undirected dependency graph is created based on the dependency tree relations.

The semantic relation extraction module adds external knowledge to the dependency graph. This knowledge may be in the form of sentiment scores stored in a lexicon or an affective resource such as SenticNet [63]. In the current study, we used sentiment scores from SenticNet as follows. For each pair of dependent words w_i and w_j in the adjacency matrix, we compute $S_{i,j}$ as:

$$S_{i,j} = Sentic(w_i) + Sentic(w_j) \tag{5}$$

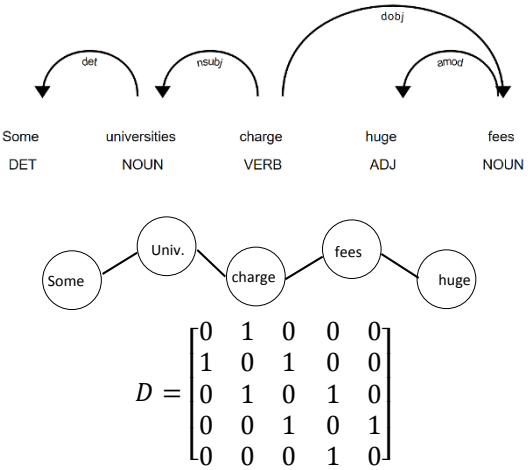


Fig. 2: A sample tweet and its corresponding dependency diagram, undirected dependency tree, and dependency graph.

where *Sentic*(w_i) is a real number in the range [-1,1] showing the sentiment intensity of w_i according to the following rule:

$$Sentic(w_i) \in \begin{cases} [-1,0), & w_i is\ a\ negative\ word \\ (0,+1], & w_i is\ a\ positive\ word \end{cases} \tag{6}$$

and if *Sentic*(w_i) = 0, w_i is a neutral word, or it is absent in SenticNet. Having computed $S_{i,j}$ for all dependent words, we enhance the adjacency matrix by:

$$A_{i,j} = D_{i,j} \times (S_{i,j} + 1) \tag{7}$$

C. GCN and Attention

The next module, GCN, takes the enhanced adjacency matrix and H^c as inputs and computes $\widetilde{H}$, which is the learned representation of the tweet as follows:

$$\widetilde{H}_i = relu(\tilde{A}_i g_i W + b) \tag{8}$$

$$g_i = \mathcal{F}(h_i) \tag{9}$$

$$\tilde{A}_i = \frac{A_i}{1 + \sum_{j=1}^{n} A_{i,j}} \quad (10)$$

where g is the hidden representation from the previous layer of GCN and $\mathcal{F}(.)$ is a transformation function, as suggested in [62].

The attention module takes H^c and $\tilde{H}$ as inputs and computes the final representation of the tweets as follows:

$$r = \sum_{i=1}^{n} \alpha_i h_i^c \quad (11)$$

$$y = softmax(W_o r + b_o) \quad (12)$$

where α_i is the attention weight calculated as follows [62]:

$$\alpha_i = \frac{e^{\beta_j}}{\sum_{j=1}^{n} e^{\beta_j}} \quad (13)$$

$$\beta_i = \sum_{j=1}^{n} h_i^{cT} \tilde{h}_i \quad (14)$$

Here, the attention mechanism is a retrieval-based method proposed by [62] and adopted in [64] for learning the affective and semantic information from a sentence.

Experimental Results

A. Datasets and Settings

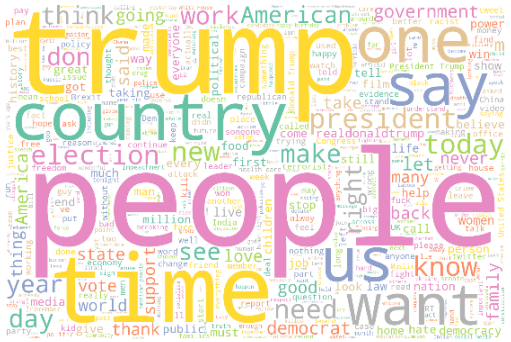
We used the RETWEET dataset [32] for experiments in the current study. The tweets in this dataset were downloaded using a pre-defined list of keywords. Word clouds of the train and test parts of RETWEET are shown in Fig. 3.

Because the public version of the RETWEET dataset only contained tweet IDs, we downloaded the tweets using the provided IDs. However, from 35020 training tweets in RETWEET, only 17613 tweets were and from 1519 test tweets, only 1037 tweets were downloaded. As discussed in the introduction section, unlike [32], we do not use the replies' textual content and only exploit the source tweets' content.

Therefore, we only need the test part of the RETWEET dataset. This dataset contains 1037 tweets, and we named it "Original". Because the Original dataset is unbalanced, we created a "Balanced" version by selecting positive, neutral, and negative tweets according to the number of tweets in the minority class (i.e., neutral class) in the Original dataset. Moreover, we created a "Resampled" version of the Original dataset by resampling the classes according to the distribution of the classes in the train part of the RETWEET dataset introduced in [32].

The detailed specifications of the datasets are shown in Table 1 and the histograms of the distribution of tweets based on their word count in the datasets are shown in Fig. 4.

In the experiments, we used the 300-dimensional vectors of Glove trained on 42 billion words from Wikipedia pages and newswires as the lookup table [58] for the proposed model. Also, in the GCN module, we used two layers, and the dimensionality of all hidden states was set to 300. The learning rate was 0.00002, the batch size was four, and the Adam optimizer with a learning rate of 0.001 was used for optimization.



(a)

(b)

Fig. 3: Word clouds of (a) the train and (b) the test parts of the RETWEET dataset.

Table 1: Specification of datasets used in the current study

	Original		Balanced		Resampled	
	train	test	train	test	train	test
Negative	318	106	226	75	167	56
Neutral	226	75	226	75	226	75
Positive	234	78	226	75	120	40
Total	778	259	678	225	513	171

B. Comparison Models

To evaluate the proposed model, the following methods were used for comparison:

- **2-BiLSTM** [32] uses two BiLSTM layers on the top of an embedding layer equipped with dropout layers.
- **2-LSTM** [65] uses two serial LSTM layers on the top of an embedding layer.

- **SenticGCN** [64] uses a GCN with depth two on the top of an LSTM layer. This model uses structural and semantic information from tweets.
- **AffectiveGCN** [64] is similar to SenticGCN but only employs semantic information to construct the dependency graph.
- **DSenticGCN** [64] similar to SenticGCN but uses directed structural graphs and GCN with depth four.

C. Evaluation Criteria

To assess the performance of models, accuracy and F1 evaluation criteria are used in the experiments.

$$F1 = \frac{2\times\pi\times\rho}{(\pi+\rho)} \quad (15)$$

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (16)$$

$$\pi = \frac{TP}{TP+FP} \quad (17)$$

$$\rho = \frac{TP}{TP+FN} \quad (18)$$

where π and ρ are precision and recall, respectively. TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively.

Results

A. Preliminary Results

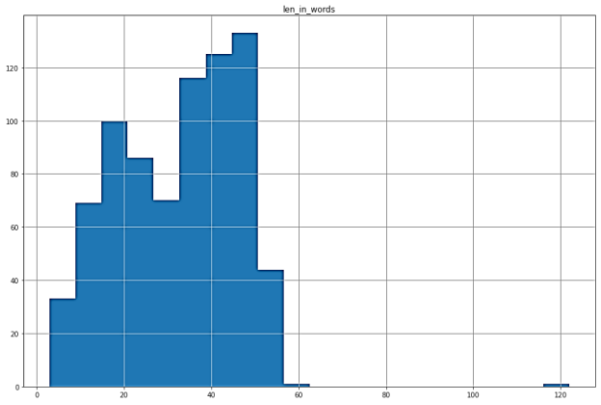
This section reports the results obtained by training the models on the ACL 2014 Twitter dataset.

As we pointed out earlier, in [32], the RETWEET model was trained on the replies posted to the original tweets (i.e., the test set of the RETWEET dataset) and evaluated on the original tweets.

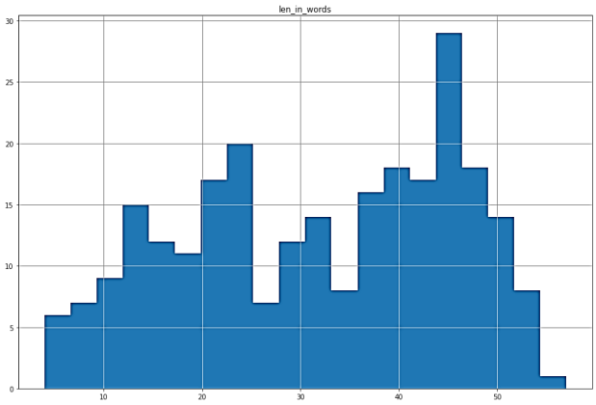
There are better methods for assessing the models than this because the final goal is to predict the overall sentiment of the first-order replies. Using these replies in the training process is unfair. Therefore, we selected the ACL 2014 Twitter dataset for the first round of experiments. This dataset contains 6248 tweets labeled as positive, neutral, or negative. The main reason for choosing this dataset for training the models is its conceptual and syntactic similarity with the RETWEET dataset.

Fig. 5 shows the results obtained using the ACL 2014 dataset for the train and test sets of three versions of the RETWEET dataset for the test.

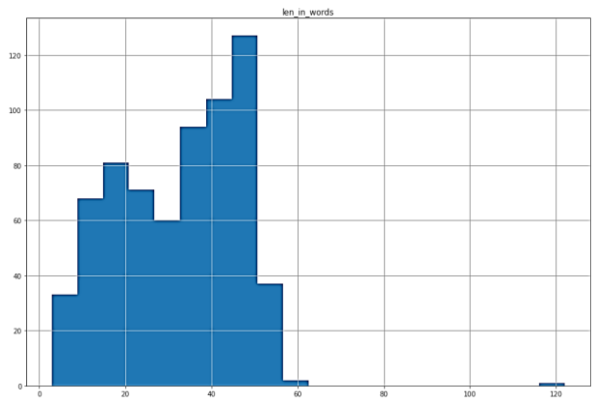
As shown in Fig. 5, when the models were trained on the ACL 2014 dataset, the proposed and other GCN-based models did not show a significant advantage. For example, on the balanced RETWEET dataset, the 2-LSTM and 2-BiLSTM models outperform other models. Also, on all test sets of the RETWEET dataset, the performance of the proposed model could have been better compared to other models when trained on the ACL 2014 dataset.



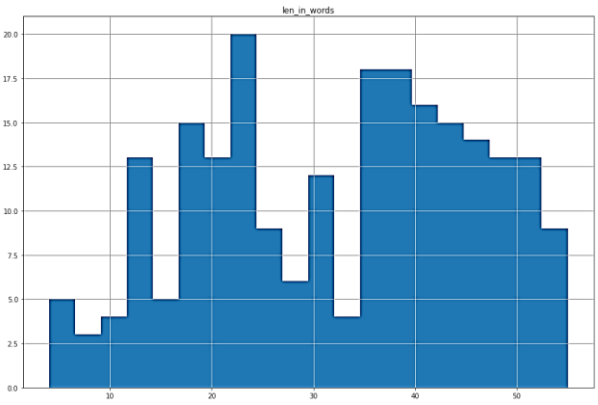
(a)



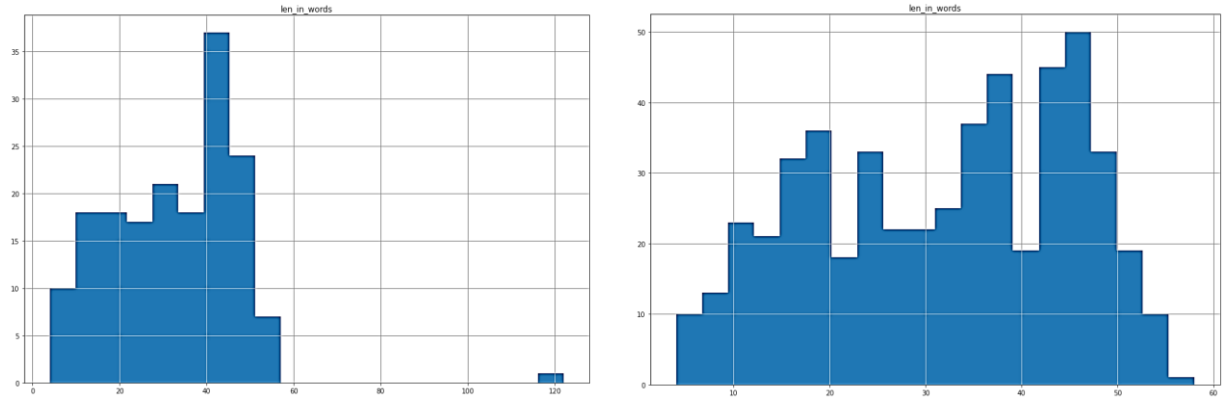
(b)



(c)



(d)



(e) (f)

Fig. 4: Histograms of the distribution of tweets based on their word count in (a) original train, (b) original test, (c) balanced train, (d) balanced test, (e) resampled train, and (f) resampled test datasets.

This lower performance of the proposed model may be justified because the proposed model's structural and semantic relation extraction modules (See Fig. 1) cannot extract meaningful relations when trained on a different dataset (i.e., the ACL 2014). To show the utility of the proposed model, in the next section, we report the results obtained using the train and test parts of the REWTWEET dataset.

B. Main Results

As shown in Fig. 6, the proposed model outperforms 2-LSTM and 2-BiLSTM models on all versions of the RETWEET dataset.

This shows the effectiveness of utilizing structural and semantic information in the proposed method. Moreover, the results show that the proposed method has similar results with other GCN-based methods on the Original RETWEET dataset while outperforming all other models on the Balanced and Resampled RETWEET datasets.

This shows the power of the proposed model for better retrieving syntactical information via the self-attention module and semantic information via external knowledge for constructing the dependency graph of the tweets. For the Balanced and Resampled datasets, the second-best method is AffectiveGCN which employs affective information in making the adjacency matrix of the tweets. This also verifies the effect of using external knowledge to enhance the system.

C. Discussion

As we pointed out in the previous section, we created three versions of the RETWEET dataset and evaluated our proposed and other deep models on these three versions. To discuss the results obtained on these three versions, we should briefly describe the process of creating the original RETWEET dataset published in [32].

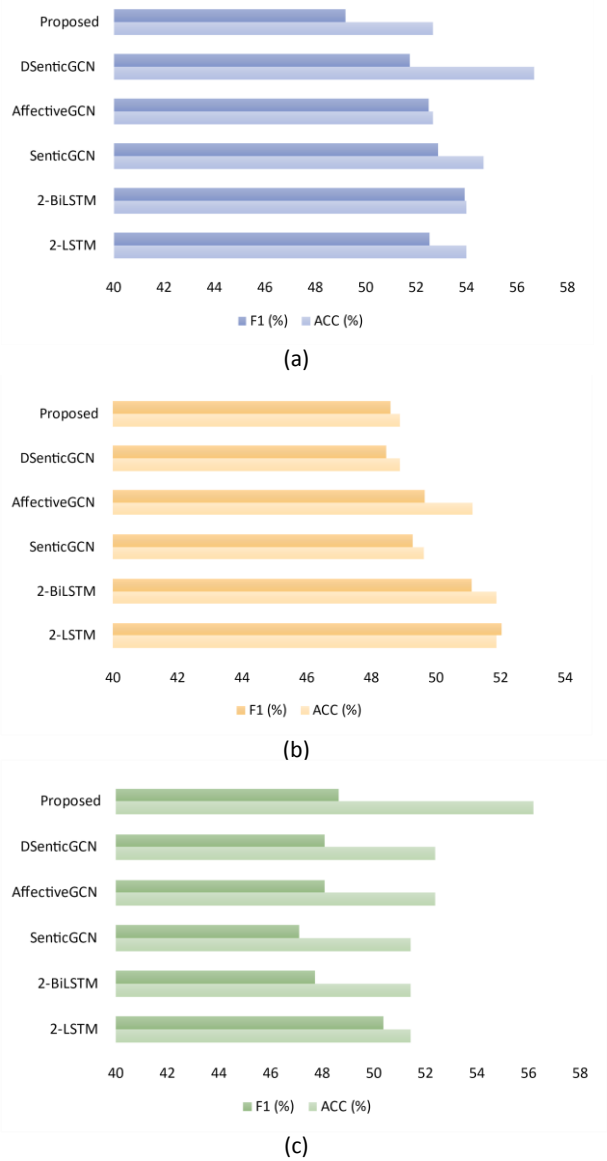
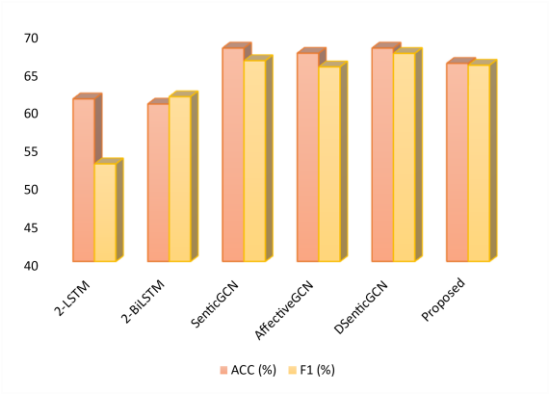
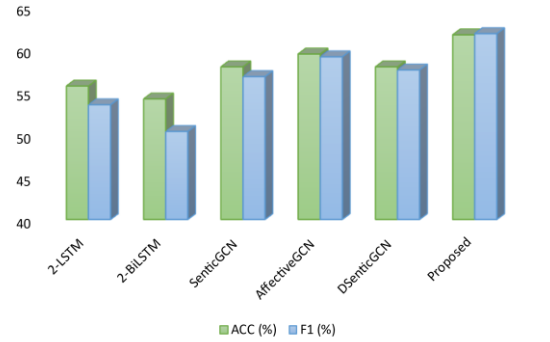


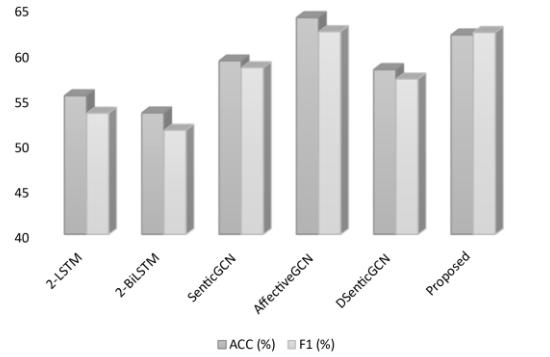
Fig. 5: Results of training the models on the ACL 2014 dataset and testing on the (a) Original, (b) Balanced, and (c) Resampled versions of the RETWEET dataset.



(a)



(b)



(c)

Fig. 6: Comparison of the results obtained by training and testing on the (a) Original, (b) Balanced, and (c) Resampled versions of the RETWEET dataset.

The first version of the RETWEET contains train and test parts created as follows [32] The training part contains first-order replies to the tweets of the test part. In the manual labeling process of the test set, three human annotators were asked to assign a three-class label to an unseen tweet based on its first-order replies in the training set. In their proposed deep model in [32], the training tweets were automatically labeled using a deep learning method (2-Bi-LSTM method in our comparisons). These labels were sent to a heuristic algorithm for assigning an overall sentiment label to the source tweet (i.e., the corresponding tweet in the test set) based on its first-order tweets. However, the original train/test separation by [32] is not helpful in the current study because we introduced a different problem in the present study. Specifically, in contrast to [32], we defined the problem as predicting the overall sentiment of replies to a tweet based on its content. Therefore, we could not use the content of the replies in the training process of the models. Hence, the original training set of [32] was useless in our study.

According to the points mentioned above, we used the original test set of RETWEET as the dataset for our experiments. We created three versions of this dataset, as described in section IV. According to this separation, the results reported in the previous section have several points which should be clarified. First, when training on another dataset (i.e., the ACL 2014 dataset), the performance of simple deep models such as 2-LSTM and 2-BiLSTM models are better (on the Balanced version) or at least comparable with more sophisticated deep models (on the Original and Resampled versions). This shows that the main reason for the effectiveness of the proposed model and similar knowledge-enhanced models is their ability to utilize the structural and semantic information in modeling the tweets in the form of an affective knowledge graph. When these models are trained on a different dataset (i.e., the ACL 2014 dataset), these models are unable to form suitable graphs and hence have weak performance.

On the other hand, when the models are trained on the RETWEET dataset, the results are more comparable and justifiable. Therefore, as the second point, it should be noted that the difference in ranking of the models in the three versions of the RETWEET dataset is due to the differences in the number of neutral, positive, and negative tweets in the datasets (i.e., see Table 1). Third, all the GCN-based models (including the proposed model) outperform the LSTM-based models, which shows the effectiveness of the graph-based convolutional models in utilizing both the tweets' textual content and their dependencies. Forth, the most confident results are for the balanced version of the dataset where the proposed model significantly outperforms all the other methods. This may be due to the simultaneous use of external knowledge and the self-attention mechanism in the proposed model.

Conclusion

Predicting the sentiment of tweet replies is an interesting problem for people and companies who want to capture other users' attention via influencing tweets. The previous studies used the replies to train deep models that can predict the tweets' overall sentiment. The content of the tweets was ignored in the process of training the deep predictive models. In the current study, we defined a new problem as follows. Given only the textual content of a source tweet, the task is to predict

the overall sentiment polarity of upcoming replies. To address this problem, we proposed a new deep model including two branches for extracting syntactical (using Bi-LSTM layers) and semantic relations (using dependency trees enhanced with an external affective source of knowledge) from the text body of the tweets and a graph convolutional network for learning the joint feature representation. Moreover, we utilized two attention mechanisms in the proposed model; first, a self-attention mechanism on the top of the Bi-LSTM module of the first branch to extract the importance of different parts of the learned representation of tweets. Second, a retrieval-based attention mechanism on the output of the graph convolutional network for learning essential features from the final affective picture of tweets.

To show the performance of the proposed model, we used the recently published RETWEET dataset, which contains manually labeled tweets in a three-class form (i.e., negative, neutral, positive) based on their content. We divided the experiments into two parts; In the first part of the experiments, we trained the models on a general tweet dataset, ACL 2014. In the second part of the experiments, we trained the models on the RETWEET dataset. The point is that when the models are trained on the RETWEET dataset, the results are more comparable and justifiable. The general ACL 2014 datasets cover several topics and users, while the RETWEET dataset contains a limited number of users and replies to their tweets. Moreover, the context of the training set is an essential factor in tuning the proposed model and similar knowledge-enhanced models because the main reason for the effectiveness of such knowledge-enhanced models is their ability to utilize the structural and semantic information in modeling the tweets in the form of affective knowledge-graph.

The results showed that the proposed model achieves similar or comparable results with simpler deep models when trained on a general tweet dataset such as ACL 2014 dataset while outperforming both simple deep models and state-of-the-art graph convolutional deep models when trained on the RETWEET dataset. This shows the proposed model's effectiveness in extracting structural and semantic relations in the tweets. For the future study, we plan to create a large dataset of tweets to address the new problem we defined in the current study. Also, designing deep ensemble models for this task may be a promising line of research for future studies.

Author Contributions

S. Nemati designed the experiments, analyzed the data, interpreted the results, and wrote the manuscript.

Acknowledgment

This work was financially supported by the research deputy of Shahrekord University (grant number 0GRD34M44264).

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work.

Abbreviations

Bi-GRU	Bidirectional Gated Recurrent Unit
Bi-LSTM	Bi-directional Long Short-Term Memory
GCN	Graph Convolutional Network
POS	Part-Of-Speech

References

[1] J. C. de Gagne, E. Cho, H. K. Park, J. D. Nam, D. Jung, "A qualitative analysis of nursing students' tweets during the COVID-19 pandemic," Nurs. Health Sci., 23(1): 273–278, 2021.

[2] M. T. Rajeh, S. N. Sembawa, A. A. Nassar, S. A. al Hebshi, K. T. Aboalshamat, M. K. Badri, "Social media as a learning tool: Dental students' perspectives," J. Dent. Educ., 85(4): 513–520, 2021.

[3] S. Rill, D. Reinel, J. Scheidt, R. V. Zicari, "Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis," Knowl. Based Syst., 69: 24–33, 2014.

[4] B. Castanho Silva, S. O. Proksch, "Politicians unleashed? Political communication on Twitter and in parliament in Western Europe," Political Sci. Res. Methods, 10(4): 776–792, 2022.

[5] N. Corsi, D. Dan Nguyen, M. Butaney, S. E. Majdalany, M. P. Corsi, T. Malchow, A. J. Piontkowski, Q. D. Trinh, S. Loeb, F. Abdollah, "Top 100 Urology Influencers on Twitter: Is Social Media Influence Associated with Academic Impact?," Eur. Urol. Focus, 9(2): 396-402, 2022.

[6] B. Mishra et al., "Use of twitter in Neurology: boon or bane?," J. Med. Internet Res., 23(5): e25229, 2021.

[7] Y. Wang, "Building relationships with fans: how sports organizations used twitter as a communication tool," Sport Soc., 24(7): 1055–1069, 2021.

[8] L. Ante, "How Elon Musk's twitter activity moves cryptocurrency markets," Technol. Forecast Soc. Change, 186: 122112, 2023.

[9] S. Duz Tan, O. Tas, "Social media sentiment in international stock returns and trading activity," J. Behav. Finance, 22(2): 221–234, 2021.

[10] E. Kouloumpis, T. Wilson, J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," in Proc. the international AAAI conference on web and social media, 5(1): 538–541, 2011.

[11] A. Giachanou, F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," ACM Comput. Surv. (CSUR), 49(2): 1–41, 2016.

[12] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, U. R. Acharrya, "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets," Knowl. Based Syst., 228: 107242, 2021.

[13] K. Sailunaz, R. Alhajj, "Emotion and sentiment analysis from Twitter text," J. Comput. Sci., 36: 101003, 2019.

[14] J. T. Yun, U. Pamuksuz, B. R. L. Duff, "Are we who we follow? Computationally analyzing human personality and brand following on Twitter," Int. J. Advert, 38(5): 776–795, 2019.

[15] M. Hasan, M. A. Orgun, R. Schwitter, "A survey on real-time event detection from the Twitter data stream," J. Inf. Sci., 44(4): 443–463, 2018.

[16] N. Oliveira, P. Cortez, N. Areal, "The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices," Expert Syst. Appl., 73: 125–144, 2017.

[17] M. H. Tsai, Y. Wang, M. Kwak, N. Rigole, "A machine learning based strategy for election result prediction," in Proc. International Conference on Computational Science and Computational Intelligence (CSCI): 1408-1410, 2019.

[18] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," Future Gener. Comput. Syst., 115: 279–294, 2021.

[19] M. Cieliebak, J. Deriu, D. Egger, F. Uzdilli, "A Twitter corpus and benchmark resources for German sentiment analysis," in Proc. 4th International Workshop on Natural Language Processing for social media: 45–55, 2017.

[20] S. Özsoy, "Use of new media by Turkish fans in sport communication: Facebook and Twitter," J. Hum. Kinet., 28: 165, 2011.

[21] S. O. Alhumoud, M. I. Altuwaijri, T. M. Albuhairi, W. M. Alohaideb, "Survey on arabic sentiment analysis in twitter," Int. J. Comput. Inf. Eng., 9(1): 364–368, 2015.

[22] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia, E. A. Villaseñor, "A case study of Spanish text transformations for twitter sentiment analysis," Expert Syst. Appl., 81: 457–471, 2017.

[23] L. Rivadeneira, J. B. Yang, M. López-Ibáñez, "Predicting tweet impact using a novel evidential reasoning prediction method," Expert Syst. Appl., 169, 2021.

[24] I. Daga, A. Gupta, R. Vardhan, P. Mukherjee, "Prediction of likes and retweets using text information retrieval," Procedia Comput. Sci., 168: 123–128, 2020.

[25] S. Butt, S. Sharma, R. Sharma, G. Sidorov, A. Gelbukh, "What goes on inside rumour and non-rumour tweets and their reactions: A psycholinguistic analyses," Comput. Human Behav., p. 107345, 2022.

[26] M. Abdar et al., "Energy choices in Alaska: Mining people's perception and attitudes from geotagged tweets," Renewable Sustainable Energy Rev., 124: 109781, 2020.

[27] S. T. Arasteh et al., "How will your Tweet be received? Predicting the sentiment polarity of Tweet replies," in Proc. 2021 IEEE 15th International Conference on Semantic Computing, ICSC 2021: 370–373, 2021.

[28] X. Gao, Z. Zheng, Q. Chu, S. Tang, G. Chen, Q. Deng, "Popularity prediction for single Tweet based on heterogeneous bass model," IEEE Trans. Knowl. Data Eng., 33(5): 2165–2178, 2021.

[29] S. Petrovic, M. Osborne, V. Lavrenko, "Rt to win! predicting message propagation in twitter," in Proc. the international AAAI conference on web and social media, 5(1): 586–589, 2021.

[30] M. Mahdikhani, "Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic," Int. J. Inf. Manage. Data Insights, 2(1): 100053, 2022.

[31] M. Pontiki et al., "SemEval-2016 task 5: Aspect based sentiment analysis," in ProWorkshop on Semantic Evaluation (SemEval-2016): 19–30, 2016.

[32] S. T. Arasteh, M. Monajem, V. Christlein, P. Heinrich, A. Nicolaou, H. Naderi Boldaji, M. Lotfinia, S. Evert, "How will your Tweet be received? Predicting the sentiment polarity of Tweet replies," in Proc. 2021 IEEE 15th International Conference on Semantic Computing, ICSC 2021: 370–373, 2021.

[33] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. J. Passonneau, "Sentiment analysis of twitter data," in Proc. the workshop on language in social media (LSM 2011): 30–38, 2011.

[34] S. M. Mohammad, S. Kiritchenko, X. D. Zhu, "NRC-Canada: Building the State-of-the-Art in sentiment analysis of Tweets," in Proc. the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), 2013.

[35] H. Ali, H. Farman, H. Yar, Z. Khan, S. Habib, A. Ammar, "Deep learning-based election results prediction using Twitter activity," Soft. Comput., 26(16): 7535–7543, 2022.

[36] L. Hong, O. Dan, B. D. Davison, "Predicting popular messages in Twitter," in Proc. the 20th international conference companion on World wide web: 57–58, 2011.

[37] Z. Luo, M. Osborne, J. Tang, T. Wang, "Who will retweet me? Finding retweeters in Twitter," in Proc. the 36th international ACM SIGIR conference on Research and development in information retrieval: 869–872, 2013.

[38] L. Wang, Y. Zhang, J. Yuan, K. Hu, S. Cao, "FEBDNN: fusion embedding-based deep neural network for user retweeting behavior prediction on social networks," Neural Comput. Appl., 34(16): 13219–13235, 2022.

[39] S. N. Firdaus, C. Ding, A. Sadeghian, "Topic specific emotion detection for retweet prediction," Int. J. Mach. Learn. Cybern., 10(8): 2071–2083, 2019.

[40] I. N. Lymperopoulos, "RC-Tweet: Modeling and predicting the popularity of tweets through the dynamics of a capacitor," Expert Syst. Appl., 163: 113785, 2021.

[41] M. D. Garvey, J. Samuel, A. Pelaez, "Would you please like my tweet?! An artificially intelligent, generative probabilistic, and econometric based system design for popularity-driven tweet content generation," Decis. Support Syst., 144: 113497, 2021.

[42] S. Poria, E. Cambria, A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," Association for Computational Linguistics, 2015.

[43] D. C. Edara, L. P. Vanukuri, V. Sistla, V. K. K. Kolli, "Sentiment analysis and text categorization of cancer medical records with LSTM," J. Ambient Intell. Hum. Comput., 14: 5309-5325, 2023.

[44] M. E. Basiri, M. Abdar, M. A. Cifci, S. Nemati, U. R. Acharya, "A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques," Knowl. Based Syst., 198: 105949, 2020.

[45] A. M. Shah, X. Yan, S. A. A. Shah, G. Mamirkulova, "Mining patient opinion to evaluate the service quality in healthcare: a deep-learning approach," J. Ambient Intell. Hum. Comput., 11: 2925-2942, 2020.

[46] M. Parimala, R. M. Swarna Priya, M. Praveen Kumar Reddy, C. Lal Chowdhary, R. Kumar Poluru, S. Khan, "Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach," Softw.: Pract. Exper., 51(3): 550-570, 2020.

[47] J. Serrano-Guerrero, M. Bani-Doumi, F. P. Romero, J. A. Olivas, "Understanding what patients think about hospitals: A deep learning approach for detecting emotions in patient opinions," Artif. Intell. Med., 128: 102298, 2022.

[48] M. E. Basiri, R. S. Chegeni, A. N. Karimvand, S. Nemati, "Bidirectional LSTM deep model for online doctor reviews polarity detection," in Proc. 2020 6th International Conference on Web Research (ICWR), 2020.

[49] H. A. Shehu et al., "Deep sentiment analysis: a case study on stemmed Turkish twitter data," IEEE Access, vol. 9, pp. 56836–56854, 2021.

[50] A. B. Nassif, A. Elnagar, I. Shahin, S. Henno, "Deep learning for arabic subjective sentiment analysis: Challenges and research opportunities," Appl. Soft. Comput., 98: 106836, 2021.

[51] H. Elfaik, E. H. Nfaoui, "Deep bidirectional lstm network learning-based sentiment analysis for arabic text," J. Intell. Syst., 30(1): 395–412, 2021.

[52] H. Saleh, S. Mostafa, L. A. Gabralla, A. O. Aseeri, S. El-Sappagh, "Enhanced arabic sentiment analysis using a novel stacking

ensemble of hybrid and deep learning models," Appl. Sci., 12(18): 8967, 2022.

[53] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, A. Hussain, "Sentiment analysis of persian movie reviews using deep learning," Entropy, 23(5): 596, 2021.

[54] Z. B. Nezhad, M. A. Deihimi, "Twitter sentiment analysis from Iran about COVID 19 vaccine," Diabetes Metab. Syndr., 16(1): 102367, 2021.

[55] S. Smetanin, M. Komarov, "Deep transfer learning baselines for sentiment analysis in Russian," Inf. Process Manag., 58(3): 102484, 2021.

[56] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[57] N. Reimers, I. Gurevych, "Alternative weighting schemes for elmo embeddings," arXiv preprint arXiv:1904.02954, 2019.

[58] J. Pennington, R. Socher, C. D. Manning, "Glove: Global vectors for word representation," in Proc. the 2014 conference on empirical methods in natural language processing (EMNLP): 1532–1543, 2014.

[59] G. Letarte, F. Paradis, P. Giguère, F. Laviolette, "Importance of self-attention for sentiment analysis," in Proc. the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP: 267–275, 2018.

[60] M. H. Phan P. O. Ogunbona, "Modelling context and syntactical features for aspect-based sentiment analysis," in Proc. the 58th Annual Meeting of the Association for Computational Linguistics: 3211–3220, 2020.

[61] Y. Vasiliev, Natural Language Processing with Python and SpaCy: A Practical Introduction. No Starch Press, 2020.

[62] C. Zhang, Q. Li, D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," arXiv preprint arXiv:1909.03477, 2019.

[63] E. Cambria, R. Speer, C. Havasi, A. Hussain, "Senticnet: A publicly available semantic resource for opinion mining.," in Proc. AAAI fall symposium: commonsense knowledge, 10(0), 2010.

[64] B. Liang, H. Su, L. Gui, E. Cambria, R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," Knowl. Based Syst., 235: 107643, 2022.

[65] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, P. Agrawal, "Understanding emotions in text using deep learning and big data," Comput. Human Behav., 93: 309–317, 2019.

Biographies

Shahla Nemati was born in Shiraz, Iran, in 1982. She received the B.S. degree in hardware engineering from Shiraz University, Shiraz, in 2005, the M.S. degree from the Isfahan University of Technology, Isfahan, Iran, in 2008, and the Ph.D. degree in computer engineering from Isfahan University, Isfahan, in 2016. Since 2017, she has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. She has written several articles in the fields of data fusion, emotion recognition, affective computing, and audio processing. Her research interests include data fusion, affective computing, and data mining.

- Email: s.nemati@sku.ac.ir
- ORCID: 0000-0003-2906-5871
- Web of Science Researcher ID: AAA-3341-2019
- Scopus Author ID: 24512475100
- Homepage: https://www.sku.ac.ir/~snemati#

How to cite this paper:
S. Nemati, "Predicting the sentiment of tweet replies using attentive graph convolutional neural networks," J. Electr. Comput. Eng. Innovations, 12(1): 57-68, 2024.

DOI: 10.22061/jecei.2023.9611.644

URL: https://jecei.sru.ac.ir/article_1926.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Uncomplicated Dead-time generation Designed for H-Bridge Drivers by Logic Gates Driving Linear Actuators

M. Karimi [*]***, D. Dideban***

Department of Electrical and Computer Engineering, University of Kashan, Kashan, Iran.

Article Info

Article History:
Received 16 April 2023
Reviewed 28 May 2023
Revised 23 June 2023
Accepted 13 August 2023

Keywords:
H-bridge
Dead-time
Shoot-through
Logic gates
Propagation delay

[*]Corresponding Author's Email Address: *Mohammadkarimi.eng@gmail.com*

Abstract

Background and Objectives: The H-bridge (HB) driver design with high efficiency is one of the most challenging issues in power systems that drive AC/DC loads. HB driver circuit based upon complementary MOSFET type used as a driving system of DC motor, power converters, and battery charger for electrical vehicles. In Driving DC motors, dead-time (DT) generation has been considered a major factor such as preventive power line short-circuits (shoot-through) over high and low-side MOSFETs. In this paper, the HB driver is designed for linear actuators with consideration for the prevention of shoot-through.
Methods: The propagation delay of logic gates are used to postpone the arrival gate drive signal for high/low side MOSFETs resulting in short circuit elimination on the DC source.
Results: As mentioned, logic gates' propagation delay by their values causes interruption between the high and low-side power switches gate drive signal resulting in shoot-through elimination. Although the existence of DT influences the performance of the rotational speed and output torque of a DC motor by increasing the distortion and pulse interval, Linear actuators due to low-velocity linear motion do not require the PWM control, therefore DT has no substantial effect on driver performance.
Conclusion: Simulation and experimental results validate the method proposed in this paper. According to the specifications of the circuit designed in this paper, for loads that do not need rotational speed control, logic gates with proper propagation delay can be chosen to eliminate short circuits in complementary MOS switches without requiring DT compensation methods.

Introduction

Linear actuators are a type of actuator that converts the rotational movement of motors into low-speed linear or straight thrust/traction motions. Linear actuators are perfect for all kinds of applications where inclination, pulling, lifting, or pushing with pounds of force is necessary, particularly in medical equipment such as Electric hospital beds or care facilities that require precision and smooth motion control. In practice, the HB driver was proposed as a power device to drive the inductive load such as a linear actuator [1].

The basic structure of HB is composed of four power transistors such as BJT, IGBT, or MOSFET. This paper uses complementary MOSFETs as power switches to drive the inductive load (Fig. 1). Due to the physical nature of power switches and body diodes, the transition delays are not the same and consequently, they cannot turn on or turn off simultaneously, IRFZ44ZSPbF (t_{on}= 14ns, t_{off}= 33ns) [2], IRF5305 (t_{on}= 14ns, t_{off} = 39ns) [3]. The intrinsic difference in timing characteristics of power switches leads to an essential issue in driving loads by the HB driver

which can be solved by DT duration "1", where t_{off} denotes the turn-off transition of one switch in the inverter leg [4].

$$t_{DT} \geq t_{off} \quad (1)$$

In practice, the difference between MOSFETs timing transitions caused short circuits among the upper and lower power switches in one inverter leg when HB was driven to change the direction of DC motor rotation or stop driving the load. To overcome this problem, an insertion delay, called DT (t_{DT}) [5], between the pulses generated by driving circuitry activates and deactivates the power switches in one branch to ensure safe operation [6], [7]. However, distortion of the output waveform, fundamental current loss [8], [9] and common-mode voltage issues [10] are the effects of DT duration, especially in the case of speed system control with high-frequency carrier and voltage source converters [11]-[13]. Therefore, DT compensation is one of the most challenging issues for the HB driver system [14], [15]. Although DT duration can prevent the breakdown of CMOS transistors in one inverter branch against shoot-through [16], it should be optimized to reduce body diode conduction and reverse recovery loss [17], [18]. In retrospective studies, different optimization methods were proposed to compensate for the DT effects in power converters and speed control systems [19]. In conventional applications, the fixed DT is used to eliminate the shoot-through [20] while the turn-off time of power switches depends on load current because of junction capacitors [21], [22]. Therefore, DT should be adjusted to an optimum value in case of varying load to eliminate shoot-through. Pulse width adjustment and adaptive DT control are the most convenient compensation methods for power electronic converters [23]-[25].

Although more compensation methods with different features were proposed, most of them require complex hardware design and precise information about the zero crossing of the load current in the existence of the noise and the current ripple which is challenging in implementation.

This paper aimed to intrinsically drive linear actuators with slow linear motion without requiring the speed control system or PWM gating signal. The approach used in this paper is based on a single pulse gate driving that the DT effect is not obvious and shoot-through elimination was guaranteed without requiring DT compensation methods. This approach leads to uncomplicated and reduced hardware utilization in HB drivers.

Modeling and Analysis of Proposed H-Bridge Driver

A. Dead-Time in H-Bridge Driver

Short circuits in complementary MOSFET or half-bridge occur when the direction of load current changes due to the gate driving signal. Changing the direction of DC motor rotation or brake state requires driving the gate of High/Low side power switches in which unequal transition delays from ON to OFF or vice versa may lead to the specific time of concurrent Complementary switches conduction resulting in DC source short circuit. In fact, in HB driver design, the gate driving control section included dead time insertion is substantial which shoot-through was eliminated. DT value depends on some characteristics of power switches such as drive current, voltage bias, input capacitance, and body diode conduction [26], [27]. Typical values for transition delays are informed in component datasheets [2], [3], so gate driving control was designed based on DT> t_{off} and referred values.

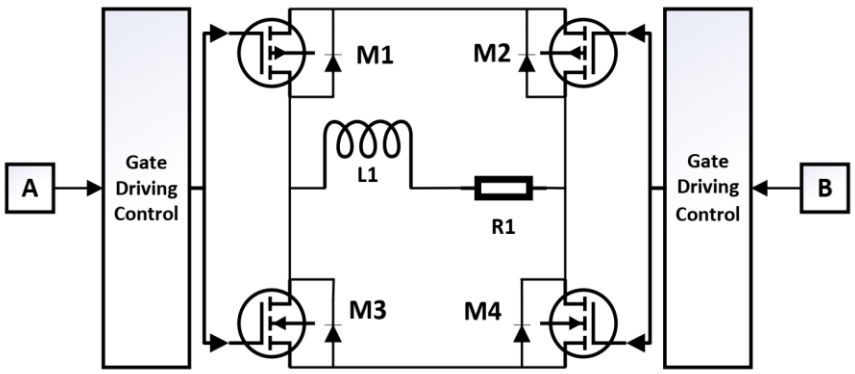


Fig. 1: CMOS H-bridge driver.

Due to the structure of the linear actuators with low-speed linear motion, speed control of the DC motor or PWM signaling is not necessary, therefore DT effects cannot degrade the efficiency of the driver. Furthermore, HB driver simulation (Fig. 2) represented the great importance of DT existence which avoids shoot-through current. The rest of this paper focused on the gate driving control part with the capability of dealing with mismatch transition delay by adding a safety time. Simulation and experimental results are provided to demonstrate the proposed method's validity and performance.

B. Proposed H-Bridge Driver Designing

Sign-magnitude (SM) and lock anti-phase are the most common driving modes in HB drivers. The proposed driver was designed based on the former. In SM mode driving, both CMOS switches are closed in each cycle and the others open.

It is necessary to add dead-time to ensure that one switch is completely off before turning on the complementary switch. In this mode during the off-time, motor winding acts as an inductor ($V_L = L \frac{di_L}{dt}$) which opposes the sudden alteration in current flowing through it. On the other hand, a sudden reduction in its current induces very high range voltages out of power switch limitations, which will destroy them. Therefore, the current should circulate in a motor rotation direction during the off-time which can be accessed by one of the turned-on switches and the other turned-off switch body diode is forward-biased. Inductor-induced voltage rose the Anode of the body diode voltage till its junction is forward-biased. Also, discrete diodes can be used along with CMOS switches instead of body diodes, but two essential characteristics in comparison to body diodes should be considered such as reverse recovery time and forward bias voltage. IRFZ44 (N-Channel) and IRF5305 (P-Channel) were chosen as complementary MOS transistors (Fig. 1). The gate driving circuitry which is comprised of fixed DT generation is incorporated into the proposed HB driver is shown in Fig. 3- 5. As depicted in the driving voltage waveforms (Fig. 4), which fed to power switches and logic gates, two direct-current (DC) potentials provide the desired biasing voltage. The first one is 24-volt which supplies the CMOS gate bias voltages by proper voltage divider resistors and the other is 5-volt which provides logic gates bias voltages. In digital signals, 5-volt generally corresponds to a high signal or '1' binary and zero potential to a low signal or '0' binary. Propagation delays of logic gates cause a blank time between activating and deactivating the CMOS transistors in one leg of the inverter. The High/Low side driver is both composed of two paths to CMOS gates. The path with NAND gates is activated during the turn-on time of the Complementary power switches (Fig. 6).

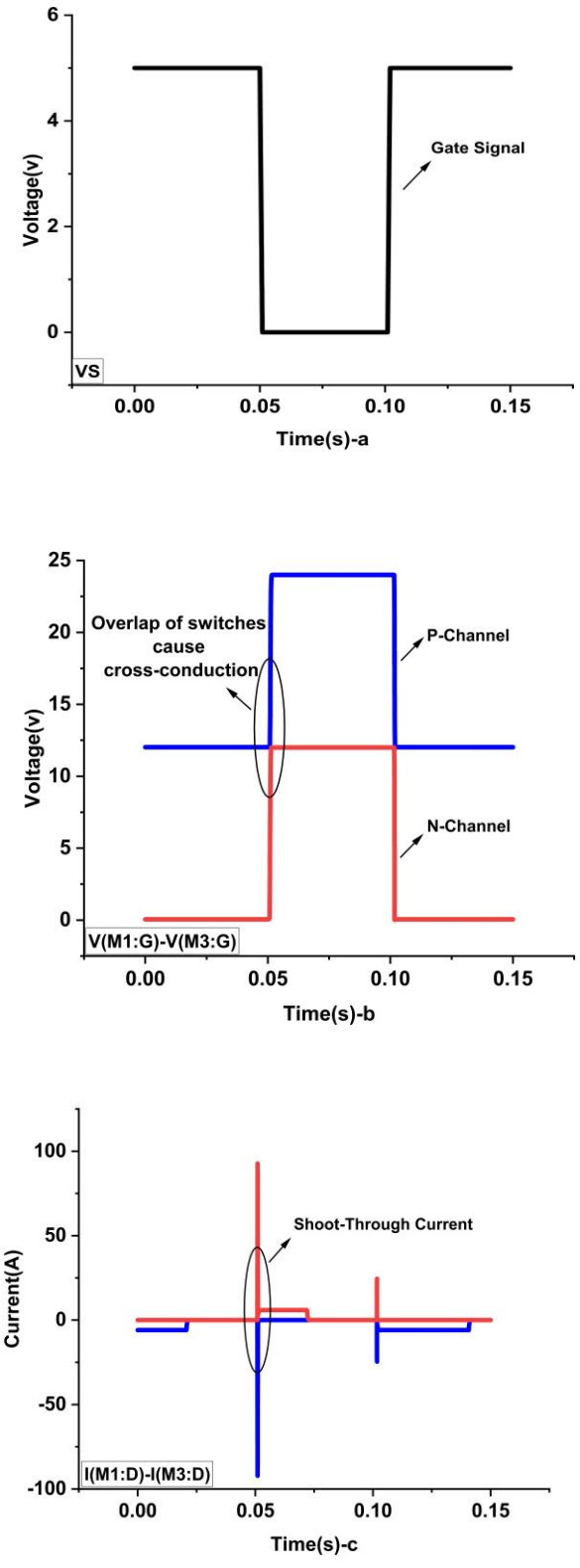


Fig. 2: Short circuit occurred during gate driving of half-bridge without DT; (a) Gate driving signal; (b) Complementary MOSFET's turn on and off concurrently; (c) Short circuit current due to both complementary power switch conduction.

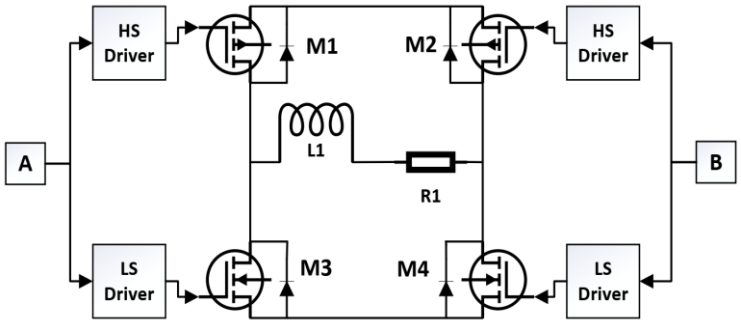


Fig. 3: Block diagram of H-Bridge circuit; HS: High-Side; LS: Low-Side.

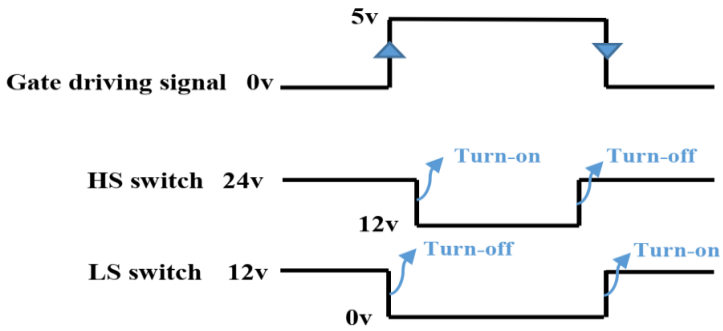


Fig.4: Gate driving signal and the sequence of activation and deactivation of High/Low side switches by the propagation delay of logic gates driving path.

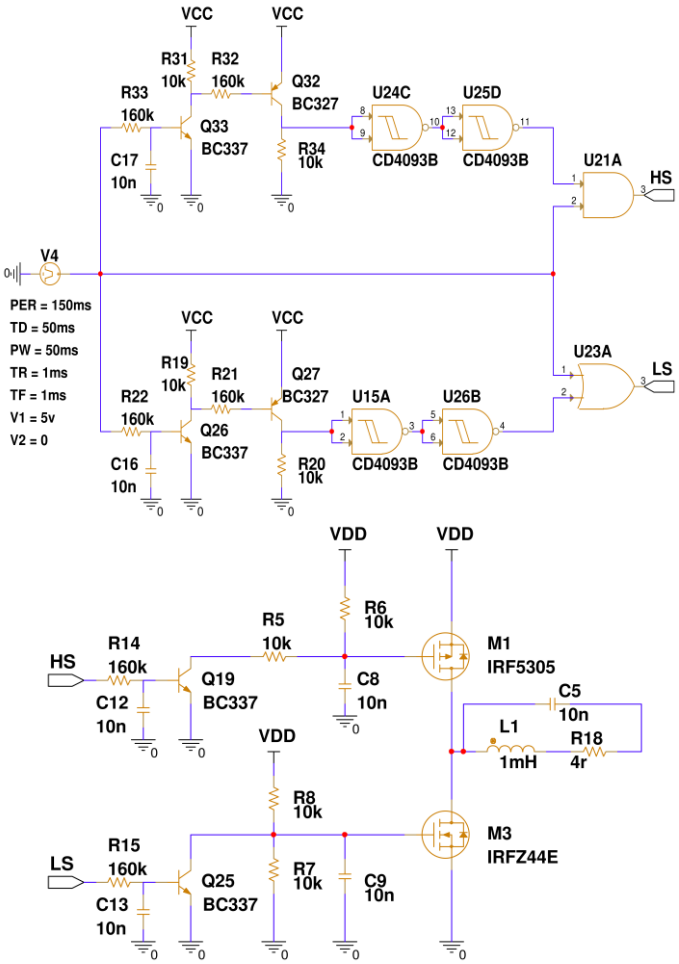


Fig. 5: Schematic diagram of proposed H-Bridge with High/Low side gate driver with NAND gates.

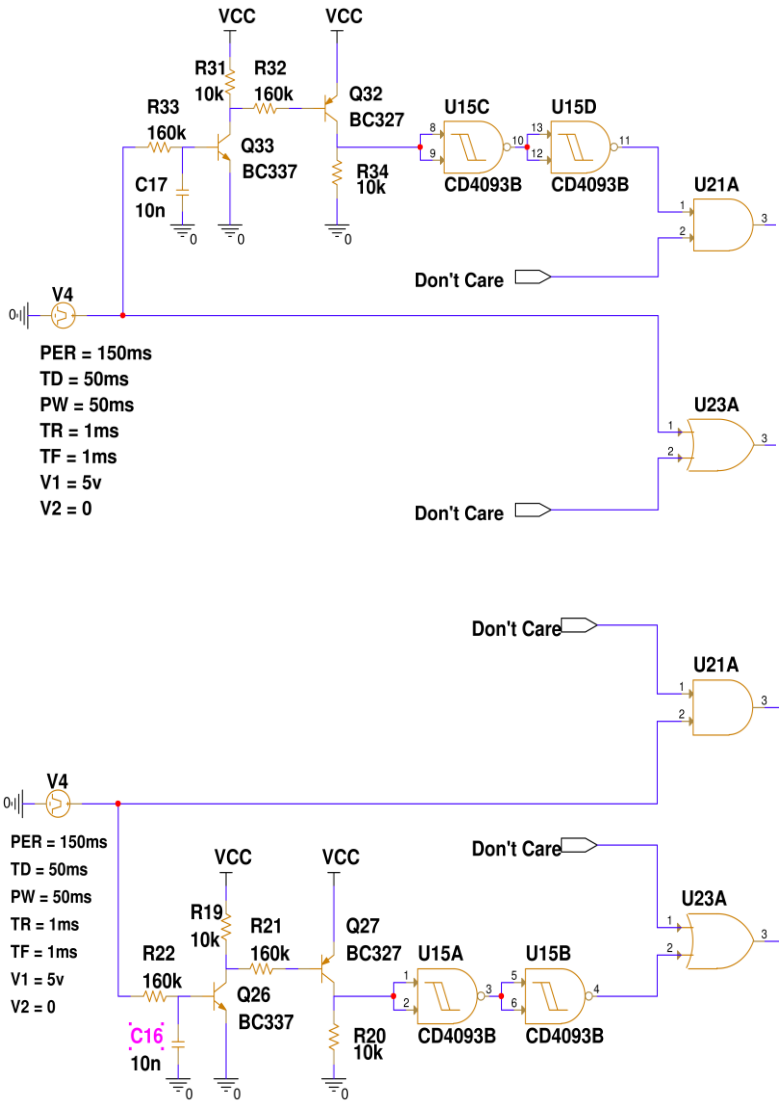


Fig. 6: NAND gates are activated in the turning-on power switch path to postpone the arrival gate drive signal.

This topology confirms one MOS transistor turns off before other complementary transistors turn on due to logic gates propagation delay (PD) acting as DT. As stated by Equ. 1, PD in whole logic gates should be upper than the CMOS turn-off time (IRFZ44: 33ns, IRF5305: 39ns). The rest of the paper is allocated to simulation and experimental results to demonstrate the validity of the proposed method.

Simulation and Experimental Results

The schematic diagram of the proposed H-Bridge in Fig. 5 was simulated in OrCAD Capture CIS version 17.2-2016. Gate driving control as shown in the schematic diagram composed of AND, NAND, and OR Logic gates in the path of high/low side arrival gates drive signal. As illustrated in Fig. 7, the outputs of logic gates level change on the length of the time interval between the specified reference points (V_M) on the input and output voltage waveforms.

These time intervals are called t_{PHL} when output switches from high to low and t_{PLH} when output switches from low to high. t_{PHL} and t_{PLH} act as DT to eliminate CMOS transistor cross-conduction. DT must be higher than the turn-off time of power switches, and the higher value of the turn-off time considered (t_{off}=39ns). According to Fig. 5, three types of logic gates were used to add proper DT in the arrival gate drive signal (Table 1).

Table 1: Types and timing characteristics of logic gates. t_{PLH}: LOW to HIGH propagation delay; t_{PHL}: HIGH to LOW propagation delay

Part Number	Type	Power Supply	t_{PLH}		t_{PHL}		Unit
			Typ	Max	Typ	Max	
HEF4093B [28]	NAND	5v	85	170	90	185	ns
HEF4081B [29]	AND	5v	45	90	55	110	ns
74LS32 [30]	OR	5v	3	11	3	11	ns

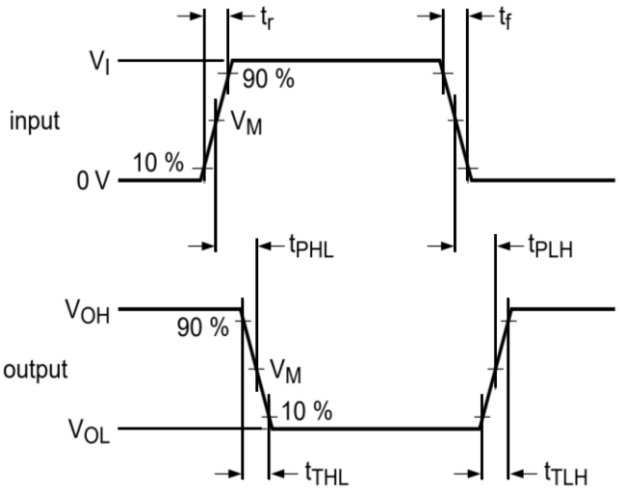


Fig. 7: Propagation delay and the output transition time of HEF4093 (NAND gate).

In accordance with different values of propagation delay for logic gate, power switches turn-on and turn-off time can be calculated. For high-side power switch (IRF5305) (Fig. 8- a):

Turn-on: t_{PLH}(NAND)+ t_{PLH}(NAND)+ t_{PLH}(AND)+ t_{on}(IRF5305) = 234 ns (2)

Turn-off: t_{PHL}(AND)+ t_{off}(IRF5305) = 94 ns (3)

For low-side power switch (IRFZ44) (Fig. 7- b):

Turn-on: t_{PLH}(NAND)+ t_{PLH}(NAND)+ t_{PHL}(OR)+ t_{on}(IRFZ44) = 192 ns (4)

Turn-off: t_{PLH}(OR)+ t_{off}(IRFZ44) = 36 ns (5)

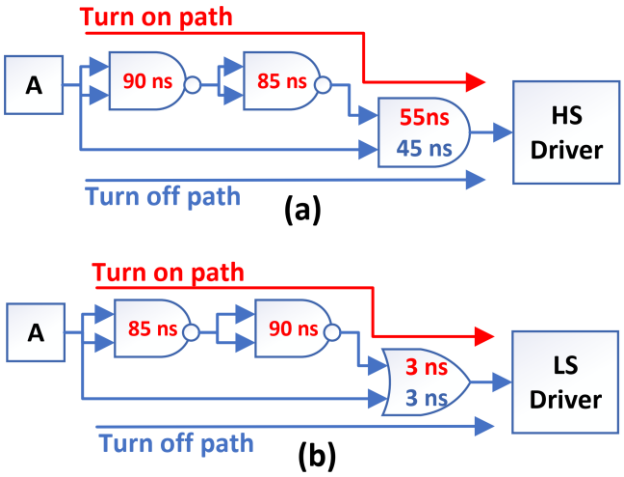


Fig. 8: High/Low side propagation delay of logic gates during turn-on or turn-off interval.

Results of calculation and simulation confirm that logic gates PD act as DT and prevent simultaneous conduction of power switches during the arrival gate drive signal (Fig. 9).

The timing diagram of logic gates proves the discrepancy between the theoretical propagation delays listed in the datasheet and practical values. As depicted in the figures, t_{PLH} and t_{PHL} are at least two orders of magnitude higher than the typical ones.

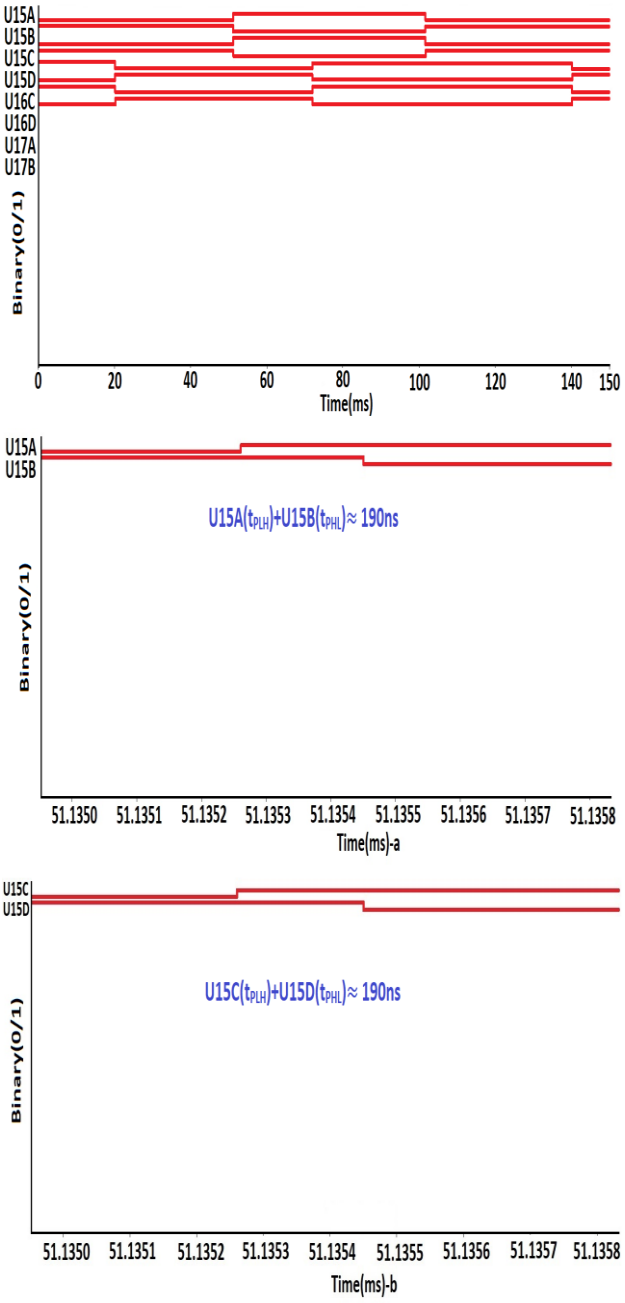
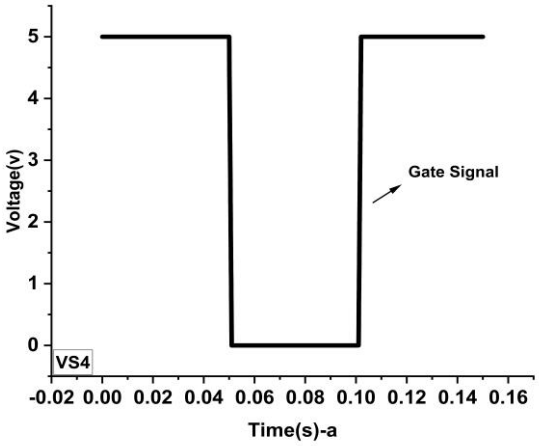
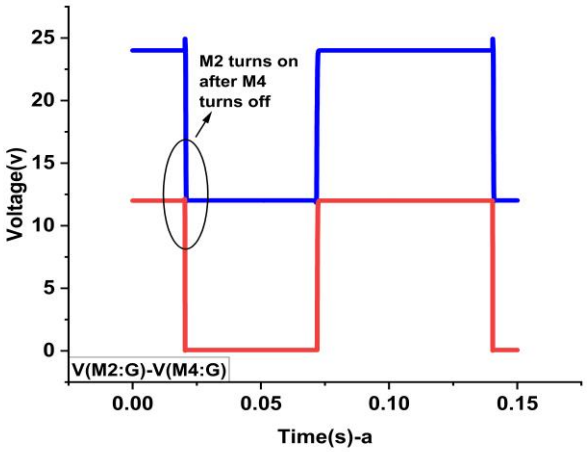
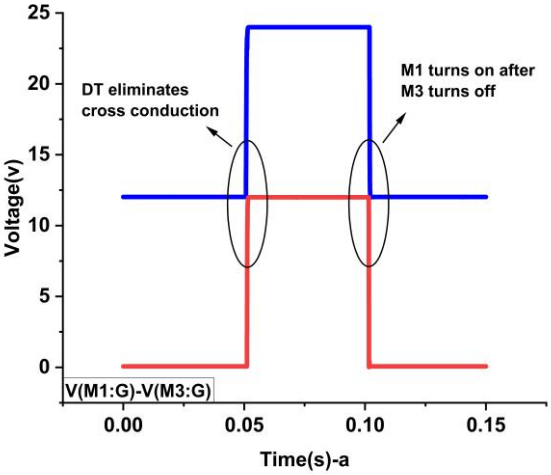


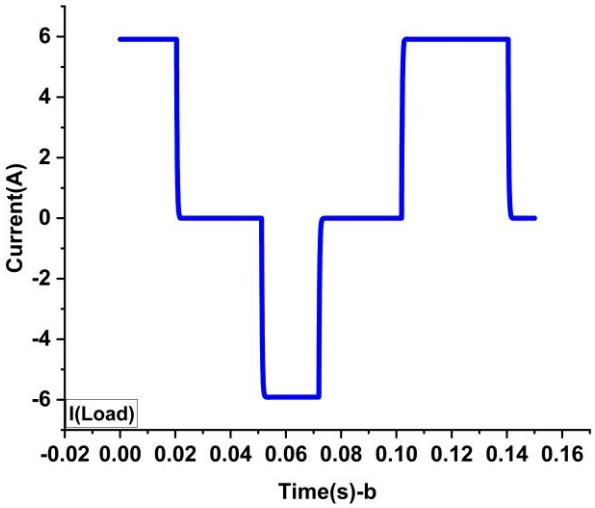
Fig. 9: Digital Timing diagram of logic gates. (a) and (b) show the propagation delay for Nand gates when different input signals impact on output (t_{PHL},t_{PLH}).

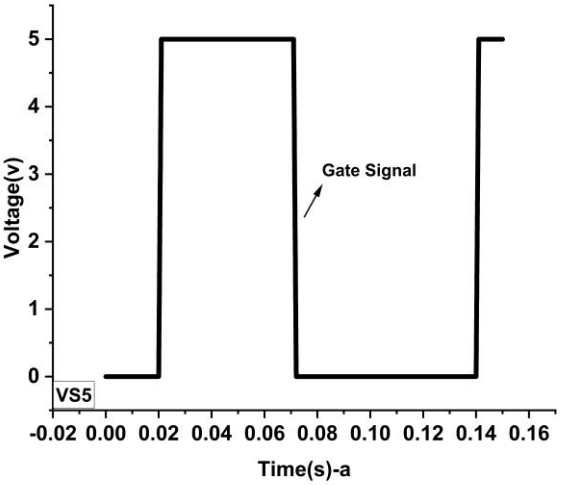
Simulation of the proposed HB driver by analyzing voltages and currents used to test the driver behavior under operating conditions that showed the validity of the schematic diagram of Fig. 5 and its method for shoot-through elimination (Fig. 10).

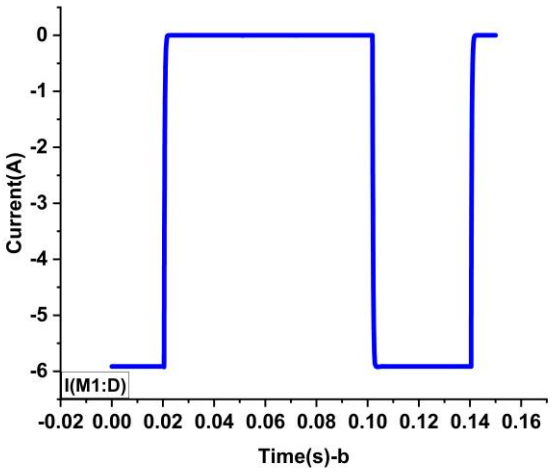

Voltage(v)
Gate Signal
VS4
Time(s)-a


M2 turns on
after M4
turns off
Voltage(v)
V(M2:G)-V(M4:G)
Time(s)-a


DT eliminates
cross conduction
M1 turns on after
M3 turns off
Voltage(v)
V(M1:G)-V(M3:G)
Time(s)-a


Current(A)
I(Load)
Time(s)-b


Voltage(v)
Gate Signal
VS5
Time(s)-a


Current(A)
I(M1:D)
Time(s)-b

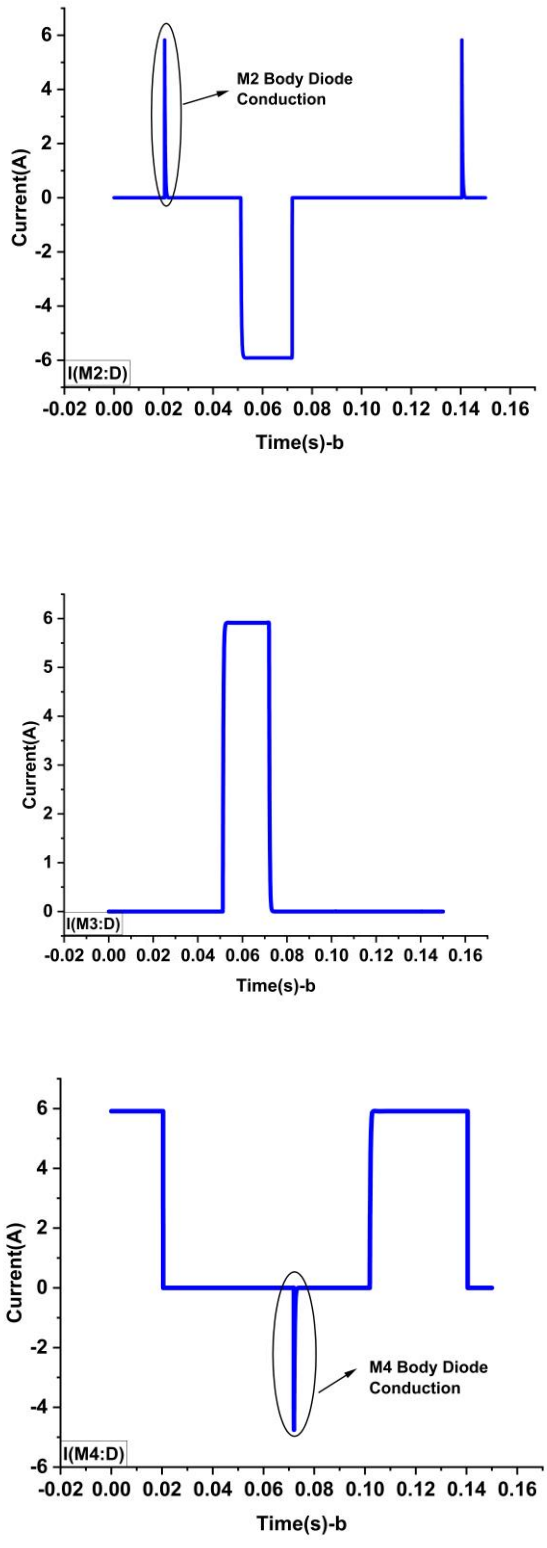


Fig. 10: voltage and current waveforms of the proposed HB driver. (a): CMOS switching voltages operation without overlapping due to the proper DT in the arrival gate drive signal; (b): load and CMOS switches currents during switching without any cross-conduction current.

Experimental results and prototype pictures of the HB driver are shown in Figs. (11-13). As expected, the proposed HB driver circuit drives the linear actuator as an inductive load without cross-conduction. Compared with other research articles [20] that fixed DT is used to overcome the cross-conduction, the method proposed in this paper is practicable and not complicated for implementing. On the other hand, most of them must use additional circuits like current polarity detection to eliminate DT effects resulting in hardware complexity and not being reliable because there is the noise and current ripple which can create a false zero crossing of the load current.

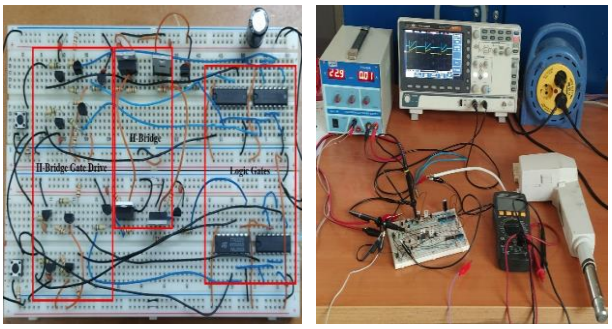


Fig. 11: prototype picture of the proposed HB driver circuit.

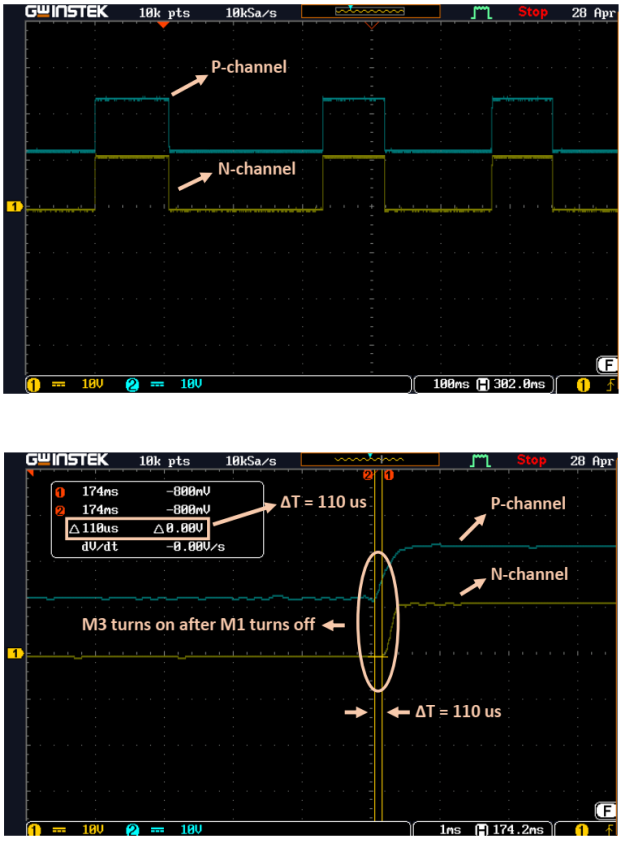


Fig. 12: Experimental results of the CMOS transistors switching while logic gates propagation delay eliminates cross-conduction without inductive load

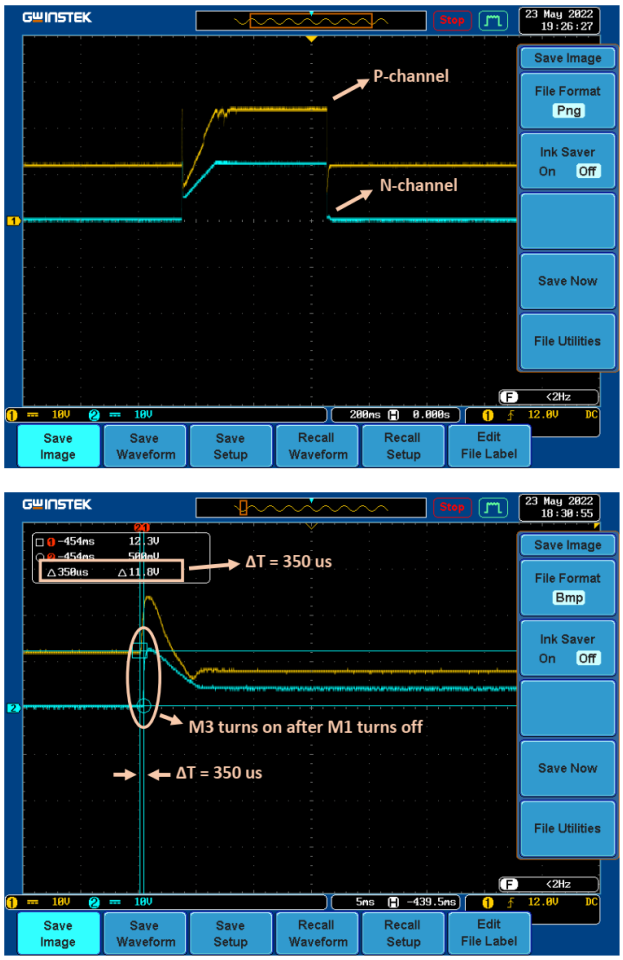


Fig. 13: Experimental results of the CMOS transistors switching while logic gates propagation delay eliminates cross-conduction with inductive load.

Conclusion

The transition delay discrepancy in MOSFET switches leads to cross-conduction in the H-bridge driver resulting in a shoot-through current. This paper investigates a novel dead-time generation method for H-bridge drivers based on CMOS transistors. As noted, dead-time should be higher than the MOSFETs turn-off time ($DT>t_{off}$) to ensure safe operation. In this paper, logic gates propagation delay included AND, NAND, and OR gates are used to generate dead-time. Dead-time value can be chosen at least two orders of magnitude higher than the turn-off time to ensure the cross-conduction elimination and the experimental results validate the accuracy of the proposed method.

Author Contributions

M. Karimi designed the experiments. M. Karimi and D. Dideban collected and carried out the data analysis. M. Karimi interpreted the results and wrote the manuscript.

Acknowledgment

Authors would like to appreciate the supports received from university of Kashan to complete this work. Moreover, various comments from anonymous reviewers helped authors to improve and enrich this research.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

DT	Dead-Time
SH	Shoot-Through
PD	Propagation Delay
DC	Direct-Current
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
CMOS	Complementary Metal Oxide Semiconductor
HS	High-Side
LS	Low-Side
t_{PLH}	LOW to HIGH propagation delay
t_{PHL}	HIGH to LOW propagation delay

References

[1] K. Mehta “Design implementation of high performance DC drives” in Proc. International Conference on Advances in Computing, Communications and Informatics (ICACCI): 740-745, 2014.

[2] International Rectifier, 55V Single N-Channel HEXFET Power MOSFET in a D2-pack package, IRFZ44ZSPbF datasheet, 2004.

[3] International Rectifier, -55V Single P-Channel HEXFET Power MOSFET in a D2-pack package, IRF5305S datasheet, 2003.

[4] Z. Zhang, F. Wang, D. J. Costinett, L. M. Tolbert, B. J. Blalock, H. Lu, "Dead-time optimization of Sic devices for voltage source converter," in Proc. 2015 IEEE Applied Power Electronics Conference and Exposition (APEC): 1145-1152, 2015.

[5] C. D. Townsend, G. Mirzaeva, G. Goodwin, “Deadtime compensation for model predictive control of power inverters,” IEEE Trans. Power Electron., 32(9): 7325-7337, 2017.

[6] P. Szczesniak, “Challenges and design requirements for industrial applications of AC/AC power converters without DC-link,” Energies, 12(8): 1581, 2019.

[7] J. Zhang, L. Fang, “An accurate approach of dead-time compensation for three-phase DC/AC inverter,” in Proc. the 2009 4th IEEE Conference on Industrial Electronics and Applications: 2929–2934, 2019.

[8] G. Liu, D. Wang, Y. Jin, M. Wang, P. Zhang, “Current-Detection-independent Dead-Time compensation method based on terminal voltage A/D conversion for PWM VSI,” IEEE Trans. Ind. Electron., 64(10): 7689–7699, 2017.

[9] E. S. Kim, U. S. Seong, J. S. Lee, S. H. Hwang, “Compensation of dead time effects in grid-tied single-phase inverter using SOGI,” in Proc. IEEE Applied Power Electronics Conference and Exposition (APEC), 2017.

[10] N. Aizawa, M. Kikuchi, H. Kubota, I. Miki, K. Matsuse, “Dead time effect and its compensation in common-mode voltage elimination

of PWM inverter with auxiliary inverter," in Proc. the 2010 International Power Electronics Conference—ECCE ASIA: 222–227, 2010.

[11] Y. Ji, Y. Yang, J. Zhou, H. Ding, X. Guo, S. Padmanaban, "Control strategies of mitigating dead-time effect on power converters: An overview," Electronics, 8(2): 196. 2019.

[12] H. Alawieh, L. Riachy, K. Arab Tehrani, K. Azzouz, B. Dakyo, "A new dead-time effect elimination method for H-bridge inverters," ECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, 2016.

[13] Z. Ming, M. Zhou, "Impact of zero-voltage notches on outputs of soft switching pulse width modulation converters," IEEE Trans. Ind. Electron., 58(6): 2345–2354, 2011.

[14] S. Iqbal, A. Xin, M. Jan, M. A. Abdelbaky, H. U. Rehman, S. Salman, M. Aurangzeb, S. Rizvi, N. Shah, "Improvement of power converters performance by an efficient use of dead time compensation technique," Appl. Sci., 10(9):3121, 2020.

[15] A. Mora, J. Juliet, A. Santander, P. Lezana, "Dead-Time and semiconductor voltage drop compensation for cascaded H-Bridge converters," IEEE Trans. Ind. Electron., 63(12):7833–7842, 2016.

[16] P. K. Chiu, P. Y. Wang, S. T. Li, C. J. Chen, Y. T. Chen, "A GaN driver IC with novel highly digitally adaptive dead-time control for synchronous rectifier buck converter," in Proc. 2020 IEEE Energy Conversion Congress and Exposition (ECCE): 3788-3792, 2020.

[17] S. Lee, S. Jung, C. Park, C. Rim, G. Cho, "Accurate dead-time control for synchronous buck converter with fast error sensing circuits," IEEE Trans. Circuits Syst. I: Regul. Pap., 60(11): 3080-3089, 2013.

[18] J. Dyer, Z. Zhang, F. Wang, D. Costinett, L. M. Tolbert, B. J. Blalock, "Online condition monitoring based dead-time compensation for high frequency SiC voltage source inverter" in Proc. 2018 IEEE Applied Power Electronics Conference and Exposition (APEC): 1854-1860, 2018.

[19] A. Lewicki, "Dead-Time effect compensation based on additional phase current measurements," IEEE Trans. Ind. Electron. 62(7): 4078-4085, 2015.

[20] B. Bellini, A. Arnaud, S. Rezk, M. Chiossi, "An integrated H-bridge circuit in a HV technology," in Proc. 2016 IEEE 7th Latin American Symposium on Circuits & Systems (LASCAS): 331-334, 2016.

[21] E. A. Jones *et al*., "Characterization of an enhancement-mode 650-V GaN HFET," in Proc. 2015 IEEE Energy Conversion Congress and Exposition(ECCE): 400-407, 2015.

[22] L. Zhang, X. Yuan, J. Zhang, X. Wu, Y. Zhang, C. Wei, "Modeling and implementation of optimal asymmetric variable dead-time setting for SiC MOSFET-Based three-phase two-level inverters," IEEE Trans. Power Electron., 34(12): 11645-11660, 2019.

[23] Y. Zhang, C. Chen, T. Liu, K. Xu, Y. Kang, H. Peng, "A high efficiency model-based adaptive dead-time control method for GaN HEMTs considering nonlinear junction capacitors in triangular current mode operation," IEEE J. Emerging Sel. Top. Power Electron., 8(1): 124-140, 2020.

[24] G. Jianning, D. Naizhe, S. Chonghui, S. Xianrui, X. Zhiwei, "Tri-carrier sinusoidal pulse-width modulation without dead time effects for converters," IET Power Electronics, 8(10): 1941-1951, 2015.

[25] L. Kang , J. Zhang, H. Zhou, Z. Zhao, X. Duan, "Model predictive current control with fixed switching frequency and dead-time compensation for single-phase PWM rectifier," Electron., 10(4): 426, 2021.

[26] Z. Zhang, F. Wang, D. J. Costinett, L. M. Tolbert, B. J. Blalock, H. Lu, "Dead-time optimization of SiC devices for voltage source converter," in Proc. 2015 IEEE Applied Power Electronics Conference and Exposition (APEC): 1145-1152, 2015.

[27] Y. Yang, Y. Tang, Y. Li, "Dead-Time elimination method of high frequency inverter with SHEPWM," in Proc. 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA): 457-461, 2019.

[28] Nexperia, Quad 2-input NAND Schmitt trigger, HEF4093B, 2015.

[29] Nexperia, Quad 2-input AND gate, HEF4081B, 2022.

[30] Texas Instruments, Quadruple 2-Input positive OR Gate, SN74LS32, 1988.

Biographies

Mohammad Karimi was born on February 1, 1991. He received M.Sc. degree in electronic engineering from Kashan University, Kashan, Iran, in 2019. He has been working as a research and development specialist in biomedical device production including physiotherapy systems, hemodialysis machines, and ventilators. His special fields of interest included Biomedical Engineering, Artificial intelligence, both analog and digital Integrated Circuits, embedded systems, and microcontroller programming.

- Email: Mohammadkarimi.eng@gmail.com
- ORCID: 0000-0003-4556-4697
- Web of Science Researcher ID: HCI-8540-2022
- Scopus Author ID: NA
- Homepage: https: NA

Daryoosh Dideban is currently an Associate professor of Nanoelectronics at the department of Electrical and Computer Engineering, University of Kashan. He received his M.Sc. and Ph.D. degrees from Sharif University of Technology, Iran, and the University of Glasgow, UK, both in Electronics Engineering. His fields of research are device simulation, compact modeling, novel two-dimensional devices, and statistical variability.

- Email: dideban@kashanu.ac.ir
- ORCID: 0000-0002-6645-1344
- Web of Science Researcher ID: AAH-6389-2020
- Scopus Author ID: 35364041200
- Homepage: https://faculty.kashanu.ac.ir/dideban/en

How to cite this paper:
M. Karimi, D. Dideban, "Uncomplicated dead-time generation designed for H-Bridge drivers by logic gates driving linear actuators," J. Electr. Comput. Eng. Innovations, 12(1): 69-78, 2024.

DOI: 10.22061/jecei.2023.9574.636

URL: https://jecei.sru.ac.ir/article_1925.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Presenting a Model of Data Anonymization in Big Data in the Context of In-Memory Processing Framework

E. Shamsinejad [1], T. Banirostam [1,], M. M. Pedram [2], A. M. Rahmani [3]*

[1]*Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran.*

[2]*Department of Electrical and Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran.*

[3]*Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.*

Article Info

Article History:
Received 19 April 2023
Reviewed 18 June 2023
Revised 05 July 2023
Accepted 15 August 2023

Keywords:
Big Data
Anonymity
Confidentiality
Data disclosure
Privacy

*Corresponding Author's Email Address:
h.banirostam.eng@iauctb.ac.ir

Abstract

Background and Objectives: Nowadays, with the rapid growth of social networks extracting valuable information from voluminous sources of social networks, alongside privacy protection and preventing the disclosure of unique data, is among the most challenging objects. In this paper, a model for maintaining privacy in big data is presented.

Methods: The proposed model is implemented with Spark in-memory tool in big data in four steps. The first step is to enter the raw data from HDFS to RDDs. The second step is to determine m clusters and cluster heads. The third step is to parallelly put the produced tuples in separate RDDs. the fourth step is to release the anonymized clusters. The suggested model is based on a K-means clustering algorithm and is located in the Spark framework. also, the proposed model uses the capacities of RDD and Mlib components. Determining the optimized cluster heads in each tuple's content, considering data type, and using the formula of the suggested solution, leads to the release of data in the optimized cluster with the lowest rate of data loss and identity disclosure.

Results: Using Spark framework Factors and Optimized Clusters in the K-means Algorithm in the proposed model, the algorithm implementation time in different megabyte intervals relies on multiple expiration time and purposeful elimination of clusters, data loss rates based on two-level clustering. According to the results of the simulations, while the volume of data increases, the rate of data loss decreases compared to FADS and FAST clustering algorithms, which is due to the increase of records in the proposed model. with the formula presented in the proposed model, how to determine the multiple selected attributes is reduced. According to the presented results and 2-anonomity, the value of the cost factor at k=9 will be at its lowest value of 0.20.

Conclusion: The proposed model provides the right balance for high-speed process execution, minimizing data loss and minimal data disclosure. Also, the mentioned model presents a parallel algorithm for increasing the efficiency in anonymizing data streams and, simultaneously, decreasing the information loss rate.

Introduction

Because sensitive data are distributed among different computational resources, in big data, unauthorized access to centralized data structures will be easily provided. The expansion of the distributed computing infrastructure, as well as the extent of mobile devices, has

raised concerns about the processing and sharing of personal and sensitive data of users [1]-[4]. In this framework, various mechanisms such as encryption, access control, audit and similar cases have been considered for maintaining data confidentiality [5]-[8].

The data stream is one of the most important big data types, which exploring them reveals hidden patterns and provides valuable information to different sciences [9]-[12]. Along with these benefits, because of the aggregation of data from various sources and exploring these data, the issue of privacy of individuals and maintaining corporate secrets are particularly regarded important. To solve this issue, various research has been done [13]. Because of their weaknesses, makes their use in big data streams impossible or suboptimal. For preventing disclosure of personal data, unique personal identifiers such as identification numbers, insurance numbers, and other distinguishing attributes are deleted before release [14]-[17]. However, after deleting the identifiers, in some cases, attackers reach personal data through public databases [18]. In order to solve this problem, a lot of research has been done to maintain the anonymity of individuals with the least changes in the dataset. In this context, methods such as k-anonymity have been proposed [19]-[21].

K-anonymity represents the anonymity by putting the tuples in K clusters. In some applications, a huge volume of data is delivered by the system in the form of a data stream that needs real time anonymization. So, anonymization of such data types through the existing algorithms is among the difficult problems. As the result, representing methods for anonymization of big data streams is inevitable [27]-[29].

In the following, the literature related to privacy protection methods in data anonymization, types of attacks and advantages and disadvantages of anonymization techniques, challenges of big data anonymization algorithms, data anonymization as well as parameters, features, methods, and algorithms of anonymization will be. Ten related works that commonly use K-anonymity for data anonymization and privacy protection for big data dissemination will be reviewed and the parameters used in the related works will be compared. And finally, the proposed model and simulations will be described and conclusions and future works will be presented.

Subject Literature

Through the data collection phase, the data publisher, who is responsible for the online anonymization of data before public release or mining, receives data streams from various sources [30], [31]. Typically, in data publishing methods, the privacy of any tuple t is considered as (1): [33].

t (Explicit Identifier, Quasi Identifier (QI), Sensitive Attributes, Non-Sensitive Attributes) (1)

The data may be streamed into the system in the form of data streams or the tables which have been stored earlier, and anonymization will be done on this kind of data. In fact, anonymity is an approach that tries to hide the identity of individuals or values of sensitive attributes from others [34].

The attacker's sum of information from public databases, her/his background knowledge, and the new anonymized database release should be less than the new database release [35]. However, it is clear that background knowledge leads to data disclosure to some extent.

Types of Privacy Methods in Data Anonymization

Privacy is one of the most important issues users confront when releasing data, especially when the data are private; such as, user's identity, location, disease background, etc.

As a result, user privacy researchers have represented various methods to protect privacy; among which, K-anonymity has been extensive.

This is true to the extent that the mentioned method is used in all environments, like centralized or distributed ones, and in services, such as centralized and distributed data mining or location-based services.

Privacy protecting methods are divided into four groups Fig. 1. Through these privacy protecting methods, all the identifiers should be deleted before release, and pseudo-identifiers, sensitive data, and non-sensitive data should be released after various anonymization operations.

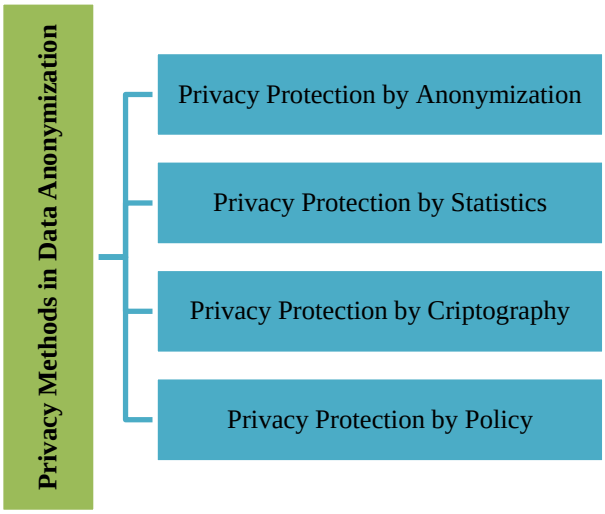


Fig. 1: Types of privacy protecting methods [59], [61], [62].

Types of Attacks

The types of attacks are divided into three categories according to Fig. 2.

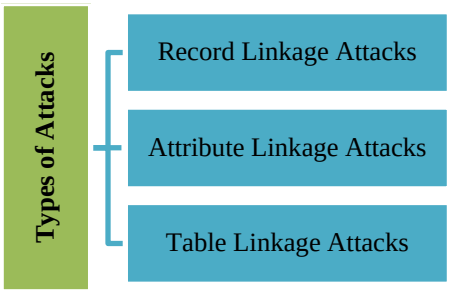


Fig. 2: Types of attacks [59].

Record Linkage Attacks

In a linkage attack record, a small number of records are distinguished based on quasi-identifier values. These numbers of records make up a group. If the quasi identifier related to the victim is mapped to this group, the attacker can identify his victim with a high probability according to his background knowledge. To deal with these types of attacks, k-anonymity was the first model offered [22]. Other models presented to contrast the record linkage attack are (x, y)-anonymity and multi-relational k-anonymity [65], [66]. These models contrast a linkage attack record by hiding the victim's report in a group with the same QI; However, if most of the reports placed in a group with the same QI have the same value for the sensitive attributes, but without accurately identifying the victim's report, the sensitive amount attributes (e.g. the type of disease) can be got. This mode is placed in the category of linkage attributes attacks [24].

Attribute Linkage Attacks

In attribute linkage attacks, the attacker may not be able to accurately determine the victim's tuple, but by mapping the victim to a group of tuple-QI with the same QI and the same amount in sensitive attributes, it can get the sensitive amount attributes of the victim with a high probability. The main idea for solving this problem is to eliminate the relationship between quasi-ID and sensitive adjective values. To solve this problem, the L-Diversity method is provided in [25]. In this method, in each group of QIs, the values of sensitive attributes should get at least l different values. In this model, if the value l is considered being k=l, k-anonymity is also guaranteed.

Other models presented to contrast attributes linkage attack are (x, y)-Privacy and (a, k)-anonymity [26] models that largely act like previous methods. If sensitive attributes are not properly distributed in the data set, the introduced models cannot contrast with the attribute linkage attack.

Suppose, in a data set, 95% of people have colds and 5% have AIDS. Now, if they have 50% AIDS and 50% cold in a QI group, a Diversity-2 condition is established. Here, the attacker can be informed of a particular person's AIDS with a 50% confidence. In the initial case, an attacker can guess a particular person's illness with a 5% confidence. T-Closeness method was presented to solve this problem in [26]. In this method, in each group of QIs, data must have a distribution close to the original data.

Table Linkage Attacks

In the attacks of the record linkage and the link of attributes, it is assumed that the attacker is aware of the existence of the victim in the published table. While sometimes the existence or absence of a person in the table can disclose sensitive information. For example, when a hospital publishes a table for AIDS patients, the knowledge of the existence or absence of a person on the table can be equal to exposing a sensitive attribute. In order to contrast this attack, a δ-present method was presented [25]. In this method, the probability of a person's presence in the published table must be limited between the two $\delta = (\delta_{min}, \delta_{max})$. This model implicitly deals with record linkage attacks and linkage attributes. In the attachment, the aggregate table compares the influential parameters of methods and algorithms. It may not imagine an end for this conflict.

Data Anonymization Techniques

In this section, a number of data anonymization techniques will be divided as shown in Fig. 3, also Table 1 will present the advantages and disadvantages of each of the introduced techniques.

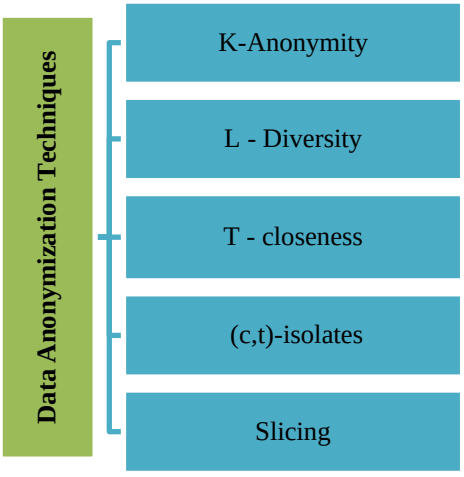


Fig. 3: Data anonymization techniques [35]-[38].

Table 1: Advantages and disadvantages of different data anonymization techniques

Anonymity Techniques	Advantage	Disadvantage
K-anonymity [23], [25]	- Simplicity in implementation - High scalability - High speed - Lower risk percentage for re-identification if K is large.	- Inefficiency against the previous knowledge of the intruder - No work against communication between data - High processing time - Inefficiency if the data is available as a query - Inefficiency in high data diversity
L–diversity [35], [36]	- Shrink and summarize data. Sensitive identifiers with equal numbers in the set. The information is repeated. - Scalability	- To be dependent on the range of changes of sensitive indicators (L-variability requires L.L is a different value for indicators.) - To be vulnerable to hacker background knowledge - In data grouping, their semantic relationship is not considered.
T-closeness [35], [36]	- Prevents skewness (sensitive diagnoses using large differences in group distribution and overall distribution).	- Computational complexity - Loss of relationship between different identifiers - low speed - Lack of scalability - Inefficiency in data diversity
(c,t)-isolation [37]	- Prevent hackers from isolating records - Scalability of performance in high data diversity	- High computational volume - Not to consider the semantic relationship between attributes - low speed
Slicing [38]	- To be suitable for high volume data - High data productivity because nothing is removed from the data.	- Due to the permutation process, it may disappear relationship between attributes. - Not to use data utility

Challenges of Anonymity in Big Data

The three main characteristics (volume, diversity and speed) provide many challenges once working with this type of data. Big data anonymization is not excluded from this rule, and in big data, in order to use any of the techniques of anonymity potential limitations and challenges must be considered. The challenges of anonymity in big data are divided into seven parts in Fig. 4.

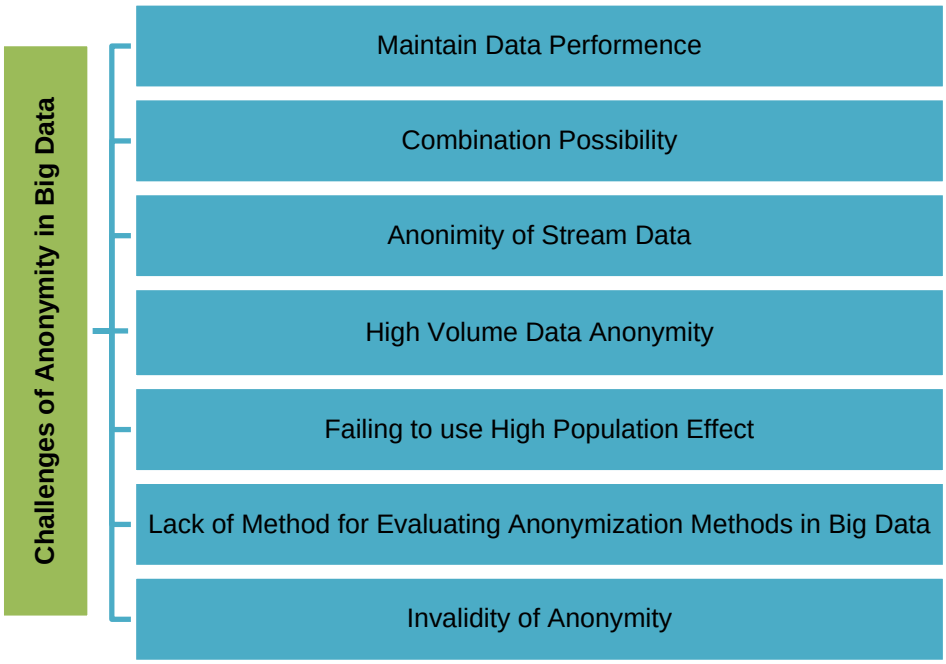


Fig. 4: Challenges of anonymity in big data [38]-[40], [63].

HDFS

Hadoop; is an open source system for distributed storage and scalable data processing. Hadoop provides a distributed file system called as Hadoop Distributed File System (HDFS) and MapReduce programming paradigm. HDFS provides to keep several copies of data and stores these copies on several nodes of cluster. It is a reliable, efficient and cost-effective system for storing large amounts of data. MapReduce provides a model for processing large amounts of data for distributed and parallel programming. The MapReduce operation basically consists of the Map and Reduce functions [18].

RDD

A new abstraction called resilient distributed datasets (RDDs) that enables efficient data reuse in a broad range of applications. RDDs are fault-tolerant, parallel data structures that let users explicitly persist intermediate results in memory, control their partitioning to optimize data placement, and manipulate them using a rich set of operators [17].

SPARK

Hadoop and Spark are two fundamental big data technologies. Hadoop provides processing data on disk while Spark process data on memory. Spark runs 100 times faster than Hadoop. This difference plays an important role for some projects requiring short response time. Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching and optimized query execution for fast queries against data of any size. Simply put, Spark is a fast and general engine for large-scale data processing. The fast part means that it's faster than previous approaches to work with Big Data like classical MapReduce. The secret for being faster is that Spark runs on memory (RAM), and that makes the processing much faster than on disk drives. Spark stores the data in the RAM of servers which allows quick access and in turn accelerates the speed of analytics [17], [18].

Anonymity Operators

Basically, datasets do not meet privacy requirements without making changes before publication. For privacy, a sequence of anonymity operators such as generalization, suppression, permutation, anatomization and perturbation are required to apply to the dataset. In the Fig. 5 shows each of the anonymous.

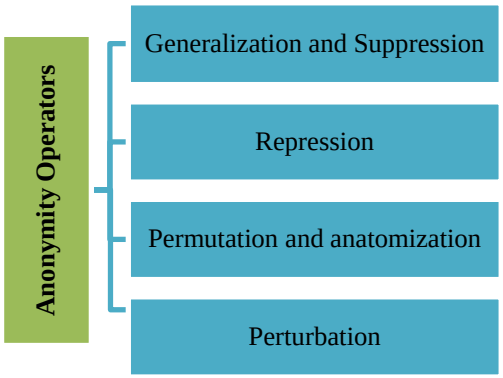


Fig. 5: Anonymity Operators [31], [33].

Anonymization Methods

Depending on the types of attack models and anonymity operators, there are various methods in data anonymization, which will be introduced in Fig. 6 and described below.

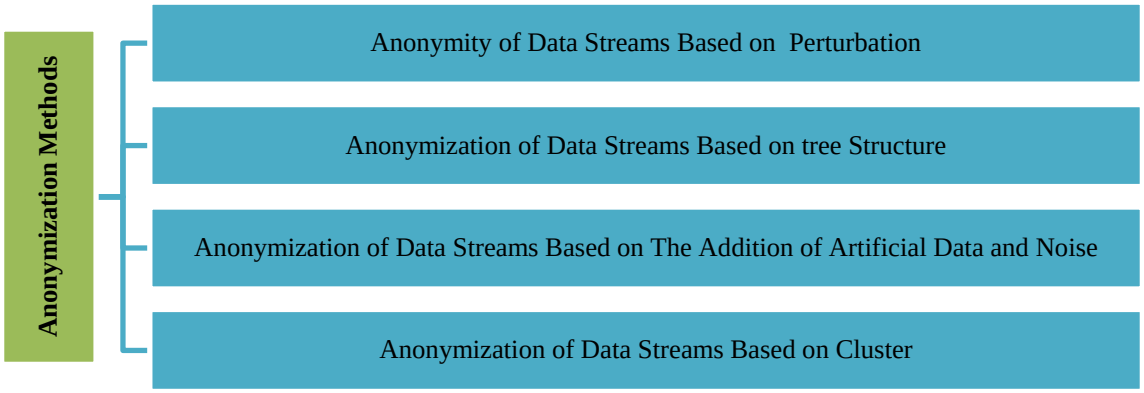


Fig. 6: Anonymization Methods [41]-[47], [60], [64].

Anonymity of Data Streams Based on Perturbation

In these algorithms, data is extracted and combined with a random noise from a statistical distribution.

The two main categories of this approach are examined below [48].

Additive Perturbation

In Additive Perturbation method, a private data set is considered as (2).

$D = d1, d2, \ldots, dn$ (2)

For each $di \in D$, random noise ri which selected from known statistical distributions such as uniform distribution and Gaussian distribution is added to the data. At last, D' dataset would be available for data miners in the form of (3) [49]-[50].

$D' = d_1+r_1, d_2+r_2, \ldots, d_n+r_n$ (3)

Data miners use $di + ri$ as a Maximal Expectation Algorithm to obtain the value di. This method of randomization is used for many data mining applications such as classification and Association Rule Mining.

Multiplicative Perturbation

One of the alternative methods proposed the Additive Perturbation method is the multiplicative Perturbation method [51]-[53]. Two common strategies in this method are derived from statistics.

In the first method, all components of D(d_i) are multiplied by a random number derived from a Gaussian distribution (usually considered one) and variance σ^2.

In the second method, the D dataset is first converted with a natural logarithm function, so that the converted components are $z_i = \ln(d_i)$. Then, a random new r_i is then added to each of the converted components, which is extracted from a multivariate equation of zero and μ Gaussian. This Gaussian distribution is considered with mean $\sigma^2 = c\Sigma z$. In this relation, $0 < c < 1$ and Σz are equal to the covariance of the converted components, that means z_i. The data that are published for data miners can be in the form of (4).

$D' = \exp(z1+r1), \exp(z2+r2), \exp(zn+rn)$ (4)

Anonymization of Data Streams Based on Tree Structure

Another category of anonymity algorithm for data streams is algorithms based on tree structures. In this context, algorithms such as SKY, SWAF and KIDS [32] have almost similar structure.

Anonymization of Data Streams Based on The Addition of Artificial Data and Noise

Despite introducing methods, in this method in order to anonymize the data stream, quasi-identifier attributes remain unchanged. In this method, data privacy is maintained by adding artificially generated data to the original data [44], [67].

Framework of Zero-Delay Anonymization Method

The main goal of this method is the real time construction of an L-variety data stream out of the main data stream. This method guarantees that the probability of guessing sensitive attributes related to a given person, in the data stream, is less than 1/1 [45].

Anonymization of Data Streams Based on Fuzzy Method

A method for protecting the privacy of the data stream is represented based on fuzzy logic [46]. In this method, the values of sensitive attributes in the data stream are converted into fuzzy values and added, in the form of a column, to the structure of records related to the same data stream.

Anonymization of Data Streams Based on Cluster

At the approach of cluster-based data anonymization, each cluster is placed in a cluster in a way that each cluster has at least tuples k. Then these tuples are published by using cluster generalization. According to Fig. 7, cluster-based anonymity algorithms will be introduced.

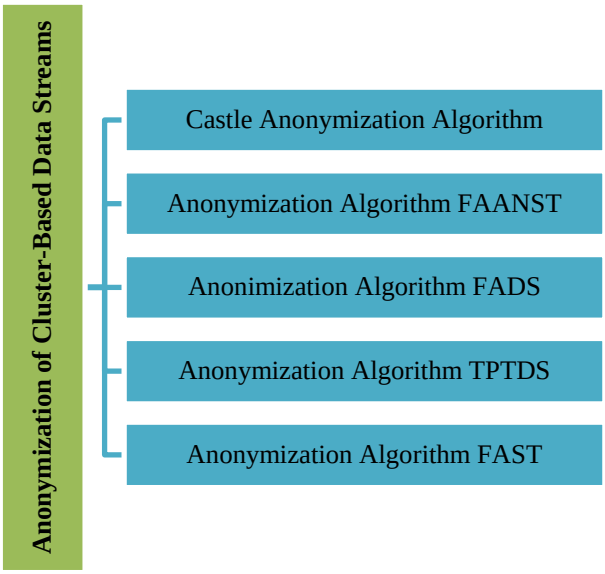


Fig. 7: Cluster-based anonymity algorithms [43]-[45].

In the following, in Table 2, the various parameters of anonymization methods and anonymization algorithms with characteristics such as data loss rate, time order, the types of data which have been subjected to anonymization approach are all examined. Furthermore, additional data production, the usability and desirability of big data technology, and the rate of response delay algorithms to access the desired data have been investigated.

Table 2: Different parameters of anonymization methods and anonymization algorithms

Algorithm	Attributes	Data loss rate	Temporal complexity	Data Type	Additional Data production	Appropriate For Big data	Delay rate
Perturbation-based	**Additive Perturbation**	Very high	O(S)	Numerical	Yes	Almost	Low
	Multiplicative Perturbation	Very high	O(S)	Numerical	Yes	Almost	Low
Tree structure-based	**SWAF**	Very high	$O(S^2 \log S)$	Numerical and deductive	No	Inappropriate	High
Artificial data-based and noise	**DF & Fuzzy**	Very high	$O\|S^2\|$	Numerical and deductive	Yes	Inappropriate	Almost low
Social Network Graphs –based [28]	**k-anonymity [28]**	Low	O((V+E)logk)	Numerical and deductive	No	Almost	High
	k-candidate [28]	Low	O((V+E)logk)	Numerical and deductive	No	Almost	High
	k-degree [28]	Low	O((V+E)logk)	Numerical and deductive	No	Almost	High
Cluster-based	**CASTEL**	Medium	$O\|S^2\|$	Numerical	No	Inappropriate	High
	FAANEST	Medium	$O\|S^2\|$	Numerical and deductive	No	Inappropriate	High
	FADS	Medium	O(S)	Numerical and deductive	No	Inappropriate	High
	TPTDS	Low	O(S)	Numerical and deductive	No	Appropriate	Not important
	FAST	Low	O(S)	Numerical and deductive	No	Appropriate	Low

Related Works

In the following, 10 related works that have been recently reviewed, will be introduced. In these works, K-anonymity method is generally used to anonymize data and maintain data confidentiality for publication. Most of these researchers have implemented their model and architecture, implementation in big data set on Adult dataset.

This dataset contains 48842 samples, 14 attributes, numerical and non-numerical data types and also 6465 missing values [50], [58].

Table 3 indicates the various parameters of the algorithms presented in the related tasks, such as data loss rate, strengths, weaknesses, time order, data types that have been anonymized and etc., are examined.

N. Victor et al. in [7].

The release of the result of a request is accompanied by noise, so the attacker might not be able of capturing information with 100% assurance. This model can protect privacy even when the attacker has background information. Even if the person does not have the correct information to publish, it has no effect on the anonymization algorithm, which is compatible with interactive and non-interactive requests.

M.Kiabod et al. in [51].

They developed an algorithm that deals with attacks in which a user's privacy is compromised by having some knowledge of a person's neighbor and friends. This algorithm tries to anonymize the data by using two techniques of generalizing data and adding additional edges to the input graph, as well as using some meta-heuristic methods. In short, this algorithm by scrolling all the input graph heads tries to expose the neighbors for the components and then isomorphic in pairs from the perspective of neighboring groups. Two important defects in this algorithm are that, firstly, a specific generalization technique has been used, which is only applicable in specific environments and specimens and does not have relative generalization. Second, during the

implementation of the algorithm to decide whether to use generalization or to add a new edge to the graph. The allocation of priority between these two actions has been used to estimate the cost of each of these two practices, which has ambiguity and seems to have a random mode.

W. Zheng et al. in [12].

Greedy algorithm is used for the anonymization of vertices' degrees, whose input and output are the graph G with n vertices and G' with anonymized k-degrees respectively. This algorithm is always capable of producing a graph of k-degrees of anonymity. The problem with this algorithm is its applicability to only one type of attack on privacy, i.e. type degree, which happens rarely compared to other types. The five methods above are investigated about the anonymization of social networks' graphs. The focus of the current article is not on this issue, but the related algorithms are discussed due to the thematic proximity.

J. Tekli et al. in [38].

Besides K-anonymity, it pays attention to the L-variety of the data. This model is based on anonymization using information clustering. Alongside receiving new data and comparing it to the existing clusters, the new data is embedded in one of the clusters if possible. In the case that the new data is not suitable for any of the clusters, the data is embedded in the most suitable cluster with the lowest rate of loss after applying the enlargement process to the clusters. The process of enlargement of a cluster is handled by extending the intervals of the respected cluster's variables. In this method, the time period, since receiving the data until releasing the data through a cluster, should not be more than a defined value σ. Regarding the threshold, it will be taken into account whether the data of unreleased clusters have reached their threshold. If such data is found, the respected cluster is released. This cluster can be immediately released when the number of data in it is greater than or equal to k; otherwise, a strategy is used to combine the cluster with the closest adjacent cluster to produce a cluster with a quantity greater than or equal to k. The writers, also, include the L-variety anonymization technique in their model so the security level increases. In this model, in a general sense, because of lack of buffering, the data are embedded in one of the existing clusters as soon as being received.

A. Otgonbayar et al. in [52].

At first, the proposed model focused on numerical data to represent a model for rapid anonymization of data streams, but it also supported non-numerical data then. Despite the represented model in CASTLE which processes and clusters the data immediately after receiving, a processing window is defined in the proposed method. Three main variables are K, MU, and DELTA which respectively refer to the anonymity variable in K-anonymity, the considered size of the processing window, and the defined threshold for information loss in each cluster. The first phase of clustering performs when the quantity of the received data in the processing window reaches MU. Some information may possibly remain in the window through this stage, not being embedded in any cluster. New data is receivable after some places in the processing window are emptied. After clustering the information, only the clusters containing data with a quantity greater than or equal to k and showing a loss rate lower than DELTA are accepted. Finally, since K-means algorithm is not usable for non-numerical attributes, clustering algorithms based on Medoid are proposed. One of the problems with this model is that no threshold is considered for preserving data in the processing window; this leads the data to remain longer in the window. This issue hinders the immediate processing of data which is among the principal necessities of data streaming**.**

J. Wang et al, in [32].

The proposed model used Encryption method to maintain confidentiality in big data. In this model, the data are first clustered using rule-based methods. Rule-based methods can be used for large volumes of data, so they can be effective for using big data. This model uses the public key asymmetric encryption method to control data access. In this model, three levels of security are considered.

- The first level, the main work of encrypting raw data is done. For this purpose, the RSA encryption method is used, which is an asymmetric encryption method. Then, the signature of the main database administrator is added to the encrypted database.
- Second level, after confirming the signature added to the database, each of the middle users ensures the accuracy of the information. They then access only part of the customer information needed for data mining and perform the desired operations. Finally, with the help of rule-based methods, information clustering operations are performed.
- The third level, which is known as the general layer, allows all users to access the extracted rules in the second level, but the original data is kept secret from users.

As mentioned, in this model, an attempt has been made to maintain the confidentiality of users' information by using encryption methods. Due to computational overhead and the need for real-time processing, encryption method is not proper for this volume of data.

B. B. Mehta et al, in [11].

Tries to provide a model for maintaining the confidentiality of big data. This model consists of three main components as follows: information anonymity

component, update component, anonymous information management component.

Anonymity operations are performed on the anonymity component of information. In this regard, the generalized method is used for anonymity of information and thus, each data is mapped to an appropriate generalized level. The update component is designed with the input of new information as well as in order to map them to the appropriate levels. After entering the information, each data is mapped to the most appropriate level of anonymity. In the meantime, with the arrival of new information, it may be necessary to make updates at the anonymity level of the database. In this case, the entire database is mapped to a new level. The anonymous information management component is responsible for maintaining the anonymous information in order to avoid the cost of recalculating the anonymous information. It can be seen that in this method, all dimensions of big data are not considered. For example, considering the "diversity" dimension in big data, this method does not provide any mechanism for assigning an appropriate level of anonymity depending on the type of input data. In this method, it should be noted that if any updates are needed, this change will be applied to the entire database. In addition to the high computational cost, this action also increases execution time, which is considered as a barrier to real-time processing. Considering the dimensions of big data, it has been tried to maintain the confidentiality of information to minimize the amount of information loss. For this purpose, an attempt is made to place the relevant data in a subgroup by dividing the data into appropriate subgroups. Due to the arrival of new information, if anonymity needs to be updated, changes need to be applied only to a portion of the database. By considering the time limit in determining the appropriate level of anonymity, the ground for real-time processing is provided.

J. Andrew et al, in [57].

The suggested model is actually an architecture based on (K, L)-anonymity. The data input source includes personal identifications from health sector, details of personal identifications, and details of individual's bank account. Some data may be received for analysis so confidentiality protection is essential before release. In the suggested architecture, first, pre-processing is done to distinguish between textual data and numerical data and classify them. The suggested architecture is designed to deal with numerical data and classifying. In the next step, anonymization techniques are carried out for the generalization of the table, through which a heuristic algorithm is used. The output of the generalized table is used as input for another confidentiality model. Then, by adding Laplacian noise, confidentiality violation is more limited. Generalization algorithm and heuristic suppression are performed based on pseudo-identifiers. First, pseudo-identifiers and sensitive identifiers are chosen according to the coefficient of their effect on confidentiality. The following criteria are considered for model evaluation: Distortion, Prec, NCP, and RMSE. The diagram comparing Distortion and Prec shows that Prec increases while K does, but after k=50 the value of Prec will be fixed. NCP evaluation results show different values of Distortion and Prec for confidentiality. The results show that the proposed method resists any type of attack.

P. Jain et al, in [67].

They represented the improved algorithms of K-Anonymization and L-Diversity for confidentiality protection in big data. They believed that these two approaches do not show hopeful results in voluminous datasets. The main issue with the current anonymization algorithms is their high rate of data loss and the huge time they need to be executed. To overcome this issue, they suggest the new models of Improved K-Anonymization (IKA) and Improved L-Diversity (ILD). IKA, through a symmetric algorithm and an asymmetric anonymization algorithm, takes K. After data anonymization using IKA, ILD is used to increase privacy. ILD makes the data more various and, consequently, increases privacy. The implementation framework for the suggested model is Apache Storm. This paper also compares the suggested model to the current anonymization algorithms like FADS, FAST, MRA, SKA. The results of implementation show that, IKA and ILD have improved significantly considering the rate of data loss and execution time.

A. Raj et al, in [66].

They presented the data anonymization algorithm with K-anonymity technique using Map-Reduce processing on a cloud base. Analyzing the data with traditional systems might be exhausting while the data volume increases. However, Map-Reduce framework is efficient and synchronizable in huge volumes of data. They presented the generalization technique for anonymization through two phases of Map and Reduce using Top Down Specification mode. Their Map-Reduce approach consists of five stages: 1-Assigning a value by Map processor to the input key K1 and sending all the data related to the mentioned key to the processor, 2-Executing each user's Map only once for each K1 and producing the gathered key values, 3- Determining the value of K2 based on the produced Map through Reduce, 4- Executing Reduce only once for each K2 produced by Map and 5- Producing the final output of Map-Reduce from all gathered and ordered Reduce outputs. The results of its application show that big data anonymization in Map-Reduce framework is efficient.

Table 3: Comparison of previous parameters and algorithms

Attributes	Researchers	Data loss rate	Strengths	Weaknesses	Chronological order	Data type	Additional data generation	Suitable for big data	Delay rate
Based on artificial data and noise	N. Victor et al, (2016) [7]	Medium	Can protect privacy even when the attacker has background information	---	---	Numerical and deductive	No	Almost	Almost low
Based on social network graphs	M.Kiabod et al, (2019) [51]	Low	---	– Using a special generalization technique – Ability to run in special environments	O((V+E)log k)	Numerical and deductive	No		High
	W. Zheng et al, (2018) [12]	Low	---	The ability to execute a special attack	O((V+E)log k	Numerical and educative	No	Almost	Almost low
Cluster based	J. Tekli et al, (2018) [39]	Medium	– Definition of the threshold for non-interruption of data release – No use of buffer – Higher security	---	---	Numerical and deductive	No	Appropriate	Almost low
	A.Otgonbayar et al, (2018) [52]	Medium	---	– No threshold is considered for preserving data in the processing window – Data remains in the window for a long time – Not suitable for real-time processing	O|S²|	Numerical	No	Appropriate	High
	J.Wang et al, (2018) [32]	Low	Suitable for big data	– Computational overhead – Not suitable for real-time processing	O|S²|	Numerical and deduct	No	Appropriate	High
	B. B. Mehta et al, (2018) [11]	Low	– Minimizing the loss rate – Being resistant to any type of attack	– Update problem – Increases execution time – Not suitable for real-time processing	O|S²|	Numerical and eductive	No	Appropriate	Almost low
	J. Andrew et al, (2020) [57]	Low	– Suitable for big data. – focus on the attributes of pseudo-identifiers. – omitting the need for previous determination of K parameter. – being needless of the awareness about the duplicated values in the columns of dataset. – being needless of adding fake data to the tables for reaching a certain K threshold. – assuring of that the highest possible value of K threshold is considered.	---	---	Numerical and deductive	No	Appropriate	High
	P.Jain et al, (2020) [67]	low	– Suitable for big data – Significant improvement in runtime and rate of data loss	---	O(n)	Numerical and deductive	No	Appropriate	Almost low
	A.Raj et al, (2019) [66]	medium	– Significant improvement in runtime and rate of data loss	---	---	Numerical and deductiv	No	Almost	Almost low
	Our Proposed model	low	– Definition of the threshold for non-interruption of data release – No use of buffer – Higher security	---	O((V+E)logk	Numerical and deductive	No	Appropriate	Almost low

Presenting the Proposed Model

To put light on the topic, the main and basic concepts used in the proposed model are represented in summary:

Methods of Determining the Optimized Quantity of Clusters

Methods of determining the optimized quantity of clusters are divided into two groups Fig. 8.

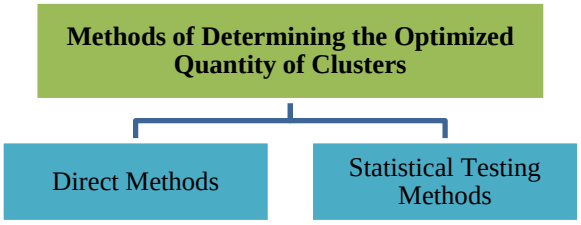


Fig. 8: Methods of determining the optimized quantity of clusters.

Direct Methods

This method seeks optimizing a particular scale like Cluster Sum of Square (WSS) or Average Silhouette. Among these methods are elbow and methods based on silhouette scale.

Statistical Testing Methods

This method seeks synchronizing the observations with a null hypothesis of a statistical test. Gap Statistics is among these methods.

The optimized quantity of clusters in K-means clustering algorithm is calculated in R programming language by the codes below (Fig. 9).

Algorithm 1

```
# Elbow method
fviz_nbclust(df, kmeans, method= "wss") + geom_vline(xintercept = 4, linetype = 2) +  labs(subtitle = "Elbow method")
# Silhouette method
fviz_nbclust(df, kmeans, method = "silhouette") +labs(subtitle = "Silhouette method")
# Gap statistic
# nboot = 50 to keep the function speedy.
# recommended value: nboot = 500 for yor analysis.
# Use verbose = FALSE to hide computing progression
set.seed(123)
fviz_nbclust(df, kmeans, nstart = 25, method ="gap_start", nboot = 50)+labs(subtitle = "Gap statistic method")
```

Fig. 9: Elbow method.

Optimal Tuple of Clusters

Most previous algorithms have used cluster generalization to send clusters. In this way, a cluster is mapped to a tuple by placing its attributes in the range obtained by an equation [54]- [56]. One of the disadvantages of this method was the high data loss rate, especially in wide-range data. In this method, we define the generalization function G, as G: PowerSet (Tuple) → Tuple with brief modifications. In this definition, Tuple represents the set of all possible tuples. PowerSet also represents all the clusters that can be defined in the S stream. The definition of the G function in (5) is fully defined.

$$G(c)=gt \quad \text{and} \quad \forall t \in c \,.\forall q \in QI \,.t \cdot q \subseteq gt \cdot q \qquad (5)$$

In this regard, QI specifies a set of identifiers. The meaning of this equation is a tuple out of a cluster is the subset of the same cluster.

Also, (6) will specify how to determine the selected tuple attributes of each cluster.

$$\begin{cases} [\text{mean}(q_1 \dots q_k) \pm \sigma] & \text{if q are numerical attribute} \\ \text{Ancestor}(q_1 \dots q_k) & \text{if q are categorical attribute} \\ \text{mod}\,(q_1 \dots q_k) & \text{if q are other attribute} \end{cases} \qquad (6)$$

In the first case, if the data type is numerical, primarily the mean and variance of the data is calculated and the interval obtained from addition and subtraction will be considered for publication. If the data type is a tree structure, the top order is selected for the existing data. If the data are different from these two cases, the data with the most repetition is considered (if the number of attributes is equal, it selects one at random). For example, considering cluster C as follows and Fig. 10, the optimal tuple of C is calculated as follows:

C={<"ali.ahmadi" , male , Academic , 42> , <"maryam.mahmoudi" , female , non_Academic , 38> , <"amin.davoodi" , male , non_Academic , 50>}

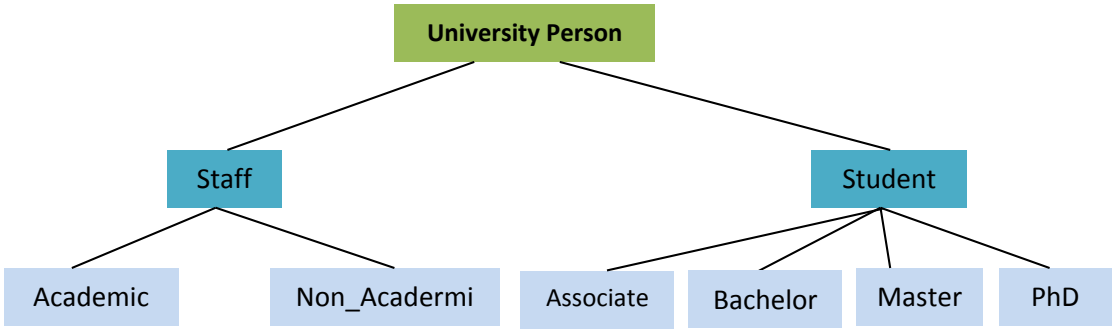


Fig. 10: University hierarchy.

G(c)=<"Explicit identifier" , mod(male, female, male), Ancestor[Academic, non_Academic], mean(42,38,50)±5]> = <*, male, staff, [38,48]>

k-anonymity Cluster

If a cluster C consists of an S stream and has more members than K, it is called a k-anonymity cluster.

Division of Clusters

Clusters whose number has reached K * k are divided into k parts by new K-means algorithm with new branches and the number of members is considered zero. K is the number of members to maintain anonymity and k is the coefficient considered in cluster division. Dividing the cluster allows new tuples to be selected in subsequent rounds, so that the values in the appropriate clusters are close together. As a result, the rate of data loss will be reduced and the result of data output analysis will be increased.

Targeted Removal of Clusters

In order not to exceed a certain number of clusters stored in the algorithm, we use (7) to determine the clusters that should be deleted. After the number of clusters exceeds the specified limit during the execution of the algorithm steps, we use the mentioned equation to remove the cluster that has the least number of members and the oldest reference time (LRU).

$$C_{del} = (1 - \alpha)n + \alpha t \quad (7)$$

In this equation, n is the number of tuples in the cluster, t is the last time a tuple refers to this cluster, and α is the coefficient that controls the weight of the two components. Clusters with the lowest values in this regard have a higher priority elimination.

Tuple Expiration Time

Real time response is among the data stream's most significant necessities which should be considered when designing anonymization systems. A tuple may remain in the system after several rounds of execution of the algorithm and will not be published and then, will be published after allowed time. This causes the system to be out of the real-time response mode and significantly increases the cost criterion. To solve this problem in the proposed algorithm, an Expiration Time (ET) parameter is considered, which indicates the maximum tolerable latency in the system. There is also a simple, innovative function called Estimated Round Time to prevent tuple publication being released after the allowed time. This function maintains the estimated time to run the next round of the algorithm and is updated in each round of running the algorithm. Next, for the remaining tuples in the system, (8) is checked and if it is correct, the tuples return to the corresponding cluster. Otherwise, the tuples should be published immediately with the cluster representative.

(Current_time- Arrival_Time)+ EstimatedRoundTime < Expiration_time (8)

Distance of Two Tuples

Suppose cluster C is to be formed of a set of data (TSet) so that |T_Set| ≥K and the amount of data loss are minimized. The closer the tuples in a data set are, the less information is lost. The distance between two tuples $t1$ and $t2$ is determined by (9).

distance (t1,t2)= w × distance(t1.QI , t2.QI) (9)

In (9), w is a vector of weight n×1, and t1. QI is a vector for tuple pseudo-identifiers t1. In fact, the weight vector of an array contains classified data that are used to determine the distance of this type of data. The distance between (t1. QI, t2. QI) is determined by the equation $distance\ (QI.\ QI,\ QI.\ QI)$ = [$d1$,..., dn]. In this equation di is calculated from (10).

$$d_i = \begin{cases} \frac{|t_l.number - t_k.number|}{t_l.number} & \text{for numerical data} \\ \frac{|t_l.level - t_k.level|}{t_l.number} & \text{for categorical data} \\ 1 & \text{if } t_l.atrrib \neq t_k.attrib \end{cases} \quad (10)$$

The method of constructing the classified data tree is based on the rules of the construction of the perfect tree in the building, and the arrangement of data is based on the properties of the tree, but the arrangement of the stream of input data in the tree formation will not be effective. Due to the need for high-speed data anonymization in real environments and the inefficiency of existing algorithms in this field and the high amount of data lost during publishing, in this section a parallel algorithm to increase efficiency in anonymizing data streams and at the same time reduction in the rate of data loss is provided. The proposed model consists of four steps. The first step is to enter the raw data from the HDFS into the RDDs, the second step is determined by a function, m clusters and their headers. In the third step, the obtained tuples are placed in parallel in separate RDDs, and finally in the fourth step, the work of classifying and publishing the clusters is done. The general process of work is shown in Fig. 11 and each step is described in detail below.

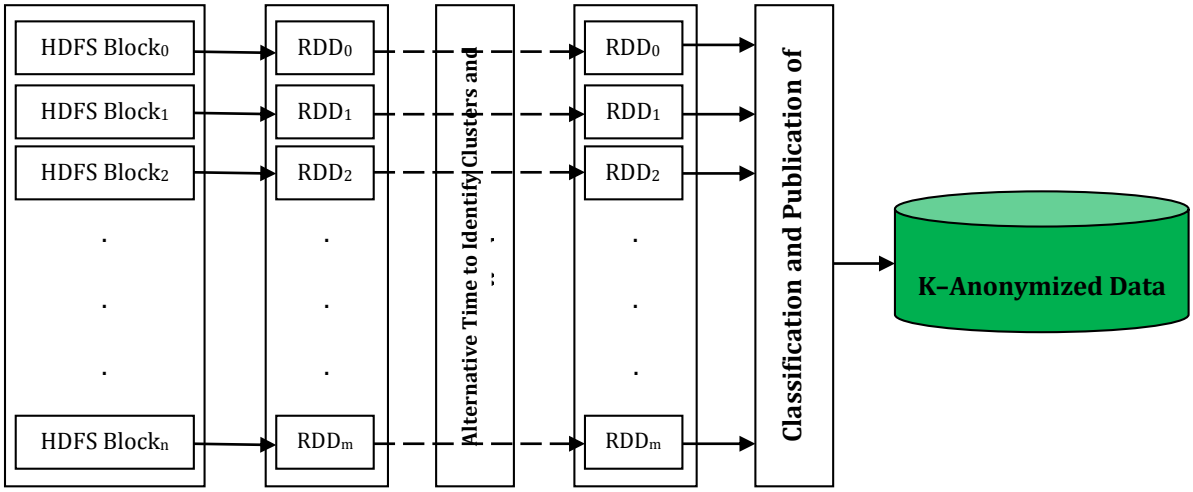


Fig. 11: The general procedure of the proposed model

Step 1. The raw data blocks in HDFS are transferred to the RDDs embedded in the system. This transfer is a type of memory transfer mapping and the number of HDFS blocks will not necessarily be equal to the number of RDDs at this stage (Fig.12).

Step 2. In this step, according to the function problem space, it is called to introduce the m point as the primary representatives of the clusters and place each in a separate cluster. This function can introduce agents randomly or by dividing the problem space. The following quasi-code describes the function of this function (Fig. 12).

Algorithm 2

```
1:  DefinitionCluster(Dataset , m)
2:    for m point do
3:      for each attribute in dataset do
4:        If (attributei was numerical ) then
5:          Split (attributei) to m segment;
6:           Insert each point of segment to agenti,m;
7:        else
8:          Random select types(attributei) without placement;
9:          Insert selected item to agenti,m;
10:       End  if
11:     End  for
12:   Insert agenti, m  into cluster Cm and add  Cm in Call;
13:  End  for
```

Fig. 12: Transfer data blocks to RDDs and introduce them points.

These clusters are then read in parallel and each placed in new RDDs. In the last step, the function that does the sorting and publishing is presented in the following quasi-code (Fig. 13).

Algorithm 3

```
1:  Main (S ,Call, Te , setexpir)
2:      Create a new thread;
3:      while S ≠ 0 do
4:          for each tuple tp in setexpir do
5:              Update expire_timetp;
6:          if (expire_timetp) > Te then
7:              Publish tp in cluster Ci whit agenti;
8:              Read tuples and inseart each  into a RDDtp;
9:          end if
10:         Call function Categorize(RDDtp, Call);
11:     End while
```

Fig. 13: Sorting and publishing.

Here S represents the inputs, Call all available clusters, Te the maximum system latency, and setexpir the unpublished tuples. The Categorize function is also defined according to the following quasi-code (Fig. 14).

Algorithm 4

```
1:  Categorize (RDDn,Call)
2:      for each agent in Call
3:          Calculate information loss between RDDn and  agenti;
4:              Insert RDDn into cluster Ci with  incures less  information loss;
5:      End for
6:     Call function PublishData(Ci,k,Nset);
7:     Terminate the thread;
```

Fig. 14: The categorize RDDs.

At this stage, by examining the size of the cluster, the necessary measures are taken for tuple publication. To preserve the K-anonymity property, minimum cluster size should be published to K, so if the number of cluster tuples reaches to K, we first calculate optimal tuple of clusters and publish the members in the cluster with the optimal calculated point. If the number of members of the cluster is greater than K and less than Nset, we publish the newly added tuple of the cluster with the optimum point calculated for the cluster in the previous steps.

A cluster whose tuple has become Nset, will first publish the newly added tuple with the optimal cluster point, then call Split Cluster function to divide the cluster. The following quasi-code checks the function of the publication function (Fig. 15).

Algorithm 5

```
1:  PublishData(Ci,k,Nset)
2:   numi=the number of members cluster Ci;
3:      If (numi ≥ k ) then
4:          If (numi = k ) then
5:              Call  function ObtimumTuple(Ci);
6:              Publish all members of the cluster Ci whit  optimum tuple calculated;
7:          End if
8:      Else if (numi = Nset ) then
9:          Publish new member of cluster Ci with optimum  tuple;
10:     Ksplit= Nset / k    \\ for K-means factor
11:  Call function SplitCluster(Call,Ci,NC, Ksplit);
12:   End if
13:  Else if (k < numi < Nset ) then
14:     Publish new member whit optimum tuple of cluster  Ci;
15:         End if
16:     Else return
```

Fig. 15: Examining size of the cluster and call SplitCluster function.

Nset is the maximum number of members allowed in each cluster. To split a cluster whose tuple of members has reached a specified number, the SplitCluster function is called. This function first checks the number of existing clusters before adding the newly formed clusters to the set and adds the clusters to the set if they are less than the allowable limit, but if the number is more than the allowed limit, it first calls the targeted deletion function of the clusters to remove the member for the new categories generated from the cluster set and then adds new clusters to the set. The members of the deleted clusters, if they have enough time, put them back in the other cluster, etc. Otherwise, it publishes it with its corresponding header. The quasi-code provides the following steps (Fig. 16).

Algorithm 6

```
1:  SplitCluster(Call,Ci,NC,Ksplit)
2:    Split Ci with Ksplit-Means algorithm;
3:    Numc = the number of members Call;
4:    If (numc + Ksplit ≥ NC) then
5:      Call  function TargetedRemove(Call, Ksplit);
6:     Add Ksplit new cluster to Call;
```

Fig. 16: Run SplitCluster.

Here, Nc refers to the maximum number of members per cluster.

So, since the proposed model is represented based on in-memory processing tools, its performance time is lower than previous models which were not implemented on a big data basis or used non-in-memory big data tools. Also, the rate of data loss decreases because of determining the non-random optimal cluster head in K-means algorithm; previous methods have used totally random determination.

Finally, the Spark logic of the model below is observed (Fig. 17).

Algorithm 7

```
Main
Input:    k-value, data, partition num, attr
Output:  k-value, node, k-Anonymized table
//Step 1: Create Taxonomy Tree and
1.        Generaliztion Lattice Tree
2.        Make_Taxonomy Tree (attr)
3.        Make_Generalization_Lattice(t-tree)
//Step 2: Make RDD and Partiton. Cache
4.        Make_RDD_From_HDFS(data);
5.        RDD_Repartition(partition num);
6.        Cache();
//Step 3: K-Anonymity (Map & Reduce)
8.        K_check = false;
9.        while(!k_check)
10.             node = next Generalization Lattice;
11.             result = MapReduce(node);
12.             k_check = Check_k_value(result);
13.       end while;
14.       Save_Output_to_HDFS(path);
```

Fig. 17: Main code.

Results and Discussion

Through the implementation of the proposed model, the dataset [58] will have various attributes, according to the Fig. 18.

```
root
 | - - CUST_ID: string  (nullable  =  true)
 | - - BALANCE: double  (nullable  =  true)
 | - - BALANCE_FREQUENCY: double  (nullable  =  true )
 | - - PURCHASES: double (nullable  =  true)
 | - - ONEOFF_PURCHASES: double (nullable  =  true)
 | - - INSTALLMENTS_PURCHASES: double (nullable  =  true)
 | - - CASH_ADVANCE: double  (nullable  =  true)
 | - - PURCHASES_FREQUENCY: double (nullable  = true)
 | - - ONEOFF_PURCHASES_FREQUENCY: double (nullable  =  true)
 | - - PURCHASES_INSTALLMENTS_FREQUENCY: double (nullable  =  true)
 | - - CASH_ADVANCE_FREQUENCY: double (nullable  =  true)
 | - - CASH_ADVANCE_TRX: integer (nullable  =  true)
 | - - PURCHASES_TRX: integer (nullable  =  true)
 | - - CREDIT_LIMUT: double (nullable  =  true)
 | - - PAYMENTS: double (nullable  =  true)
 | - - MINIMUM_PAYMENTS: double  (nullable  =  true)
 | - - PRC_FULL_PAYMENT: double  (nullable  =  true)
 | - - TENURE: integer (nullable =  true)
```

Fig. 18: Various attributes dataset [58].

First, preprocessing and standardization of the given data will be exerted according to the code below (Fig. 19).

```
from pyspark.ml.feature import VectorAssembler
data_customer.columns
assemble=VectorAssembler(inputCols=[
  ' BALANCE ',
  ' BALANCE_FREQUENCY ',
  ' PURCHASES ',
  ' ONEOFF_PURCHASES ',
  ' INSTALLMENTS_PURCHASES ' ,
  ' CASH_ADVANCE ',
  ' PURCHASES_FREQUENCY ',
  ' ONEOFF_PURCHASES_FREQUENCY ',
  ' PURCHASES_INSTALLMENTS_FREQUENCY ',
  ' CASH_ADVANCE_FREQUENCY ',
  ' CASH_ADVANCE_TRX ',
  ' PURCHASES_TRX ' ,
  ' CREDIT_LIMIT ',
  ' PAYMENTS ',
  ' MINIMUM_PAYMENTS ' ,
  ' PRC_FULL_PAYMENT ' ,
  ' TENURE ' ], outputCol=' features ')
assembled_data=assemble.transform(data_customer)
assembled_data.show (2)
```

Fig. 19: Preprocessing codes.

The results of preprocessing the first two lines are illustrated in Fig. 20.

```
+-------+-----------+-----------------+---------+-------------------+--------------------+-----------+-------------------+-----
----------------------+-----------------------------------+-----------------------+-----------+------------+-----------+-----
------+----------------+-----------------+-------+--------------------+
|CUST_ID|    BALANCE|BALANCE_FREQUENCY|PURCHASES|ONEOFF_PURCHASES|INSTALLMENTS_PURCHASES|CASH_ADVANCE|PURCHASES_FREQUENCY|ONEOF
F_PURCHASES_FREQUENCY|PURCHASES_INSTALLMENTS_FREQUENCY|CASH_ADVANCE_FREQUENCY|CASH_ADVANCE_TRX|PURCHASES_TRX|CREDIT_LIMIT|   PA
YMENTS|MINIMUM_PAYMENTS|PRC_FULL_PAYMENT|TENURE|            features|
+-------+-----------+-----------------+---------+-------------------+--------------------+-----------+-------------------+-----
----------------------+-----------------------------------+-----------------------+-----------+------------+-----------+-----
------+----------------+-----------------+-------+--------------------+
| C10001| 40.900749|         0.818182|     95.4|                0.0|                95.4|        0.0|          0.166667|
0.0|         0.083333|              0.0|        0|          2|     1000.0| 201.802084|      139.5
09787|            0.0|   12|[40.900749,0.8181...|
| C10002|3202.467416|         0.909091|      0.0|                0.0|                0.0| 6442.945483|               0.0|
0.0|              0.0|             0.25|        4|          0|     7000.0|4103.032597|     1072.3
40217|         0.222222|   12|(17,[0,1,5,9,10,1...|
+-------+-----------+-----------------+---------+-------------------+--------------------+-----------+-------------------+-----
----------------------+-----------------------------------+-----------------------+-----------+------------+-----------+-----
------+----------------+-----------------+-------+--------------------+
```

Fig. 20: The results of preprocessing the first two lines.

When implementing the proposed method, k=2 has been considered; the results are illustrated in Table 4.

Table 4: Results of implementing the proposed model with k=2

Original Table			
RID	**Age**	**Gender**	**Disease**
1	31	M(1)	Diabetes
2	21	M(1)	Anemia
3	26	F(0)	Pneumonia
4	36	F(0)	Anemia
5	34	M(1)	Diabetes
6	25	F(0)	Pneumonia

2-Anonyized Table			
RID	**Age**	**Gender**	**Disease**
1	**30~39**	***(0~1)**	**Diabetes**
2	**20~29**	***(0~1)**	**Anemia**
3	**20~29**	***(0~1)**	**Pneumonia**
4	**30~39**	***(0~1)**	**Anemia**
5	**30~39**	***(0~1)**	**Diabetes**
6	**20~29**	***(0~1)**	**Pneumonia**

According to Fig. 21, the results of the loss criterion have reached the stability value of 0.30.

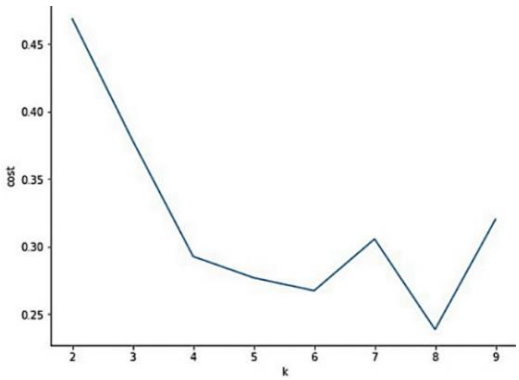


Fig. 21: Results of the loss criterion with k=9.

Finally, the rate of data loss in the suggested model with two other clustering algorithms, i.e. FADS and FAST, are shown with different values of K for 10 and 100 megabytes of data are shown in Figs. 22 and 23 respectively.

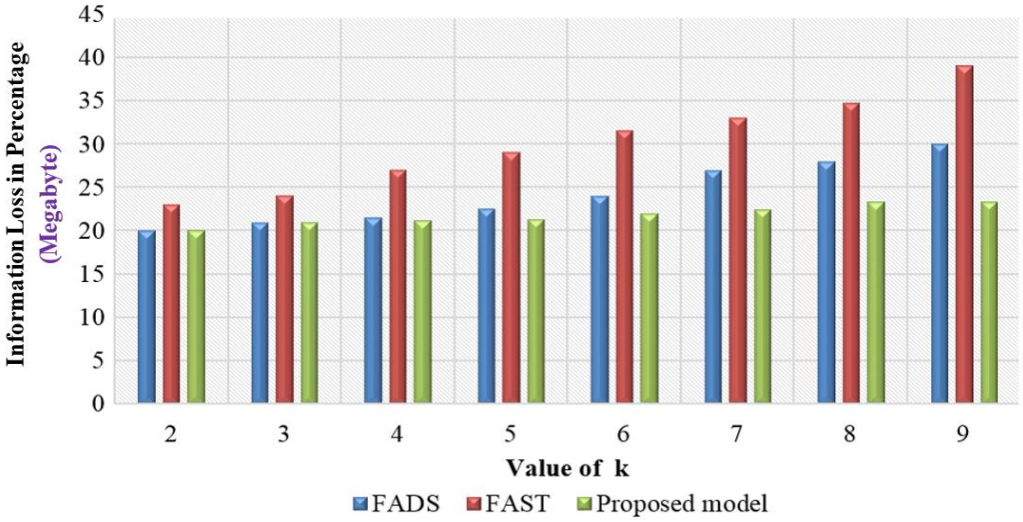


Fig. 22: The rate of data loss in FADS and FAST clustering algorithms and the suggested model with 10 megabytes of data.

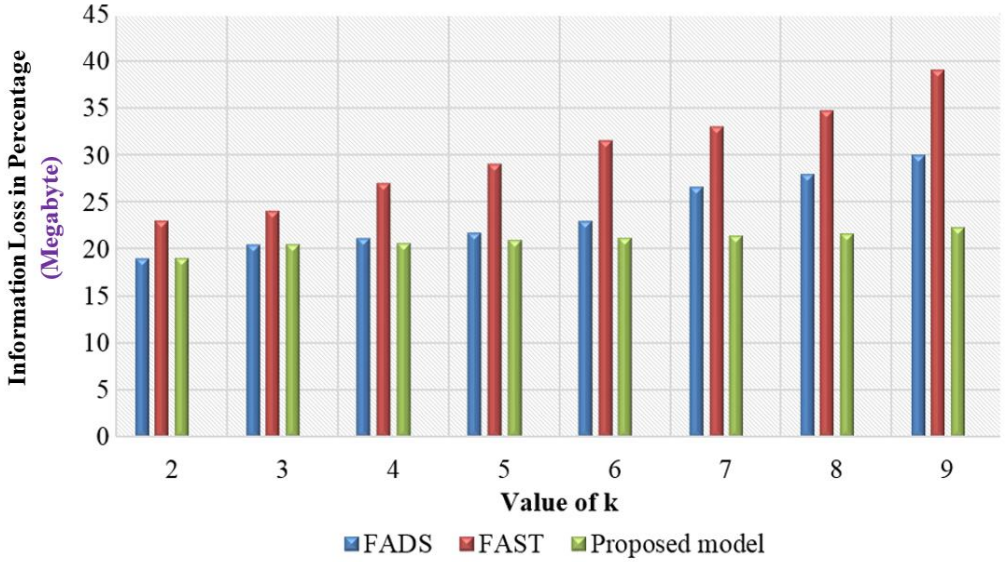


Fig. 23: The rate of data loss in FADS and FAST clustering algorithms and the suggested model with 100 megabytes of data.

Considering the results of simulation in Figs. 23 and 24 the rate of data loss in suggested model, in both cases of 10 megabytes and 100 megabytes of data, is lower than that in FAST and FADS clustering algorithms so, despite the increase of data volume, the rate of data loss is reduced because of increasing the records.

Conclusions and Future Works

Anonymization is not limited only to omitting some attributes and replacing some values with other ones; actually, it is an effort for finding a method to make an optimized relation between privacy protection and the possibility of using data while decreasing the rate of information loss. In this paper, various methods of anonymization, such as anonymization based on perturbation, based on a tree structure, based on zero-delay, based on the addition of artificial data, based on fuzzy method, based on clustering, and common algorithms are presented, accompanied by a comparison of their various parameters. The architectures and models in the related literature which have dealt with big data anonymization are investigated, and factors such as the data loss rate, amount of extra data production, suitability for big data environment, and delay time are compared. Reviewing the related literature, it is revealed that there still exists the necessity of high-speed data anonymization in real environments, the inefficiency of the current algorithms, and the high rate of data loss through release. However, the results of investigating the suggested model and solution show that clustering by in-memory processing of Spark platform provides a suitable and reasonable time for the anonymized release of big data, and the designed steps reduce the information loss to the lowest possible amount. Considering the results, while the data volume increases, the rate of data loss decreases compared to FADS and FAST clustering algorithms which are because of increasing the records in the suggested model.

Author Contributions

All the authors participated in all aspects of the preparation and writing of this article.

Acknowledgment

The authors thankfully appreciate the anonymous reviewers and the editor of JECEI for their useful comments and suggestions.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
ET	Expiration Time
FAANST	Fast Anonymizing Algorithm for Numerical Streaming data
FADS	Feature anomaly detection system
HDFS	Hadoop Distributed File System
IKA	Improved K-Anonymization
ILD	Improved L-Diversity
KIDS	K-anonymization Data Stream
LRU	Least Recently Used
Mllib	Machine Learning Library
MRA	MapReduce based Anonymization
NCP	Normalized Certainty Penalty
RDD	Resilient Distributed Datasets
RMSE	Root Mean Square Error
RSA	Rivest Shamir Adleman
SKA	Scalable k-Anonymization
SKY	Stream K-anonymity
SWAF	Sliding Window Anonymization Framework
TKC	Tanzu Kubernetes Cluster
TPTDS	Two Phase Top-Down Specialization
WSS	Within Cluster Sums of Squares

References

[1] P. Zhao, H. Jiang, C. Wang, H. Huang, G. Liu, Y. Yang, "On the performance of k-anonymity against inference attacks with background information," IEEE Internet Things J., 6(1): 808-819, 2019.

[2] S. Sangeetha, G. Sudha Sadasivam, Handbook of Big Data and IOT Security, first ed., Springer, Switzerland, 2019.

[3] S. Patnaik, New Paradigm of Industry 4.0: Internet of Things, Big Data & Cyber Physical Systems, first ed., Springer, Switzerland, 2019.

[4] A. Chaudhary, Ch. Choudhary, M. Kumar Gupta, Ch. Lal, T. Badal, Microservices in Big Data Analytics, first ed., Springer, Singapore, 2019.

[5] X. Zhang, Ch. Liu, S. Nepal, Ch. Yang, J. Chen, Security, Privacy and Trust in Cloud Systems, first ed., Springer, Berlin, 2013.

[6] J. Salas, J. Domingo-Ferrer, "Some basics on privacy techniques, anonymization and their big data challenges," Math. Comput. Sci., 12: 263–274, 2018.

[7] N. Victor, D. Lopez, "Privacy models for big data: A survey," J. Big Data Intel., 3: 61-75, 2016.

[8] K-K. Raymond Choo, A. Dehghantanha, Handbook of Big Data Privacy, Springer, Switzerland, 2020.

[9] M. Al-Zobbi, S. Shahrestani, Ch. Ruan, "Improving mapreduce privacy by implementing multi-dimensional sensitivity-based anonymization", J. Big Data., 4(1): 1-23, 2017.

[10] Sh. Luan Hou, X. Kun Huang, Ch. Qun Fei, Sh. Han Zhang, Y. Yang Li, Q. Lin Sun, Ch. Qing Wang, "A survey of text summarization approaches based on deep learning," J. Comput. Sci. Technol., 36: 633-663, 2021.

[11] B. B Mehta, P. Rao U, "Toward scalable anonymization for privacy-preserving big data publishing," Adv. Intel. Syst. Comput., 2: 297-304, 2018.

[12] W. Zheng, Z. Wang, T. Lv, Y. Ma, C. Jia, "K-Anonymity algorithm based on improved clustering," ICA3PP, 11335: 462-476, 2018.

[13] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, McCauley, M. J. Franklin, S. Shenker, I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," NSDI'12, 15-28, 2012.

[14] P. Ram Mohan Rao, S. Murali Krishna, A. P. Siva Kumar, "Privacy preservation techniques in big data analytics: A survey," J. Big Data., 5: 1-12, 2018.

[15] S. Khan, Kh. Iqbal, S. Faizullah, M. Fahad, J. Ali, W. Ahmed, "Clustering based privacy preserving of big data using fuzzification and anonymization operation," Int. J. Adv. Comput. Sci. Appl. (IJACSA), 10(12): 282-289, 2019.

[16] A. Dobson, K. Roy, X. Yuan, J. Xu, "Performance Evaluation of machine learning algorithms in apache spark for intrusion detection," in Proc. International Telecommunication Networks and Applications Conference (ITNAC), 127:1-6, 2018.

[17] S. Ullah Bazai, J. Jang-Jaccard, "SparkDA: RDD-Based high-performance data anonymization technique for spark platform," in Proc. International Conference on Network and System Security, 11928: 646-662, 2019.

[18] Y. Canbay, S. Sagiroglu, "Big data anonymization with spark, in Proc. International Conference on Computer Science and Engineering, (UBMK): 833-838, 2017.

[19] M. Al-Zobbi, S. Shahrestani, Ch. Ruan, "Experimenting sensitivity-based anonymization framework in apache spark," J. Big Data., 5: 1-26, 2018.

[20] M. Mittal, V. E. Balas, L. Mohan Goyal, R.Kumar, Big Data Processing Using Spark in Cloud, first ed., Springer, Singapore, 2019.

[21] Z. He, H. Cai, "Latent-Data privacy preserving with customized data utility for social network data," IEEE Trans. Veh. Technol., 67(1): 665-673, 2018.

[22] B. Matturdi, X. Zhou, S. Li, F. Lin "Big data security and privacy: a review," China Commun., 11(14): 135-145, 2014.

[23] Z. Ouazzani, H. Bakkali, "A new technique ensuring privacy in big data: k-anonymity without prior value of the threshold k," Procedia Comput. Sci., 127: 52-59, 2018.

[24] F. Fei, S. Li, H. Dai, C. Hu, W. Dou, Q. Ni, "A k-anonymity based schema for location privacy preservation," IEEE Trans. Sustainable Comput., 4(2): 156-167, 2019.

[25] Y. Canbay, Y. Vural, S. Sagiroglu, "Privacy Preserving Big Data," in Proc. International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), 24-29, 2018.

[26] A. Kayem, C. T. Vester, Ch. Meinel, "Automated k-anonymization and l-diversity for shared data privacy," in Proc. International Conference on Database and Expert Systems Applications (DEXA), 9827: 105-120, 2016.

[27] J. Shish Patel, S. Priyanka, "Online analytical processing for business intelligence in big data," J. Big Data, 8(6): 501-518, 2020.

[28] K. R. Macwan, S. J. Patel, "k-NMF anonymization in social network data publishing," Secur. Comput. Syst. Networks Comput., 61(4): 601–613, 2018.

[29] A. Reiza, M. A. Armengol de la Hoz, M. S. Garcíaa, "Big data analysis and machine learning in intensive care units," Med. Intensiva, 43(7): 416-426, 2019.

[30] J. Novotny, P. A. Bilokon, A. Galiotos, F. Deleze, Machine Learning and Big Data with kdb+/q, first ed., Wiley, London, 2020.

[31] M. Bowles, Machine Learning with Spark and Python, Second ed., John Wiley & Sons., Indianapolis, 2020.

[32] J. Wang., Zh. Cai, Y. Li, D. Yang, L. Li, H. Gao, "Protecting query privacy with differentially private k-anonymityin location-based services," Pers. Ubiquitous Comput., 22: 453–469, 2018.

[33] L. Arbuckle, Kh. El Emam, Building an Anonymization Pipeline, first ed., O'Reilly Media, California, 2020.

[34] S. Ram Prasad Reddy, K. V.S.V.N. Raju, V. Valli Kumari, "Personalized privacy preserving incremental data dissemination through optimal generalization," J. Eng. Appl. Sci., 13(11): 4205–4216, 2018.

[35] J. Domingo-Ferrer, "Big data anonymization requirements vs privacy models," in Proc. International Conference on E-Business and Telecommunication Networks (ICETE), 2: 305-312, 2018.

[36] S. A Abdelhameed, Sh. M Moussa, M. E Khalifa, "Restricted sensitive attributes-based sequential anonymization (RSA-SA) approach for privacy-preserving data stream publishing," Knowledge-Based Syst., 164: 1-20, 2019.

[37] Y. Canbay, A. Kalyoncu, M. Ercimen, A. Dogan, S. Sagiroglu, "A clustering based anonymization model for big data," in Proc. International Conference on Computer Science and Engineering (UBMK): 720-725, 2019.

[38] J. Tekli, B. Al Bouna, Y. Bou Issa, M. Kamradt, R. Haraty, "(k, l)-clustering for transactional data streams anonymization," International Conference on Information Security Practice and Experience (ISPEC), 11125: 544-556, 2018.

[39] P. Jain, M. Gyanchandani, N. Khare, "Improved k-anonymity privacy-preserving algorithm using madhya pradesh state election commission big data," Commun. Security, Stud. Comput. Intel., 771: 1-10, 2019.

[40] K. Guo, Q. Zhang, "Fast clustering-based anonymization approaches with time constraints for data streams," J. Software, 24: 1852-1867, 2014.

[41] Y. Wang, Zh. Chi, X. Tong, L. Li, "A differentially k-anonymity-based location privacy-preserving for mobile crowdsourcing systems," Procedia Comput. Sci., 129: 28-34, 2018.

[42] C. Eyupoglu, M. Aydin, A. Zaim, A. Sertbas, "An Efficient big data anonymization algorithm based on chaos and perturbation techniques," Entropy, 20(5): 1-18, 2018.

[43] A. Nezarat, Kh. Yavari, "A distributed method based on mondrian algorithm for big data anonymization," in Proc. International Congress on High-Performance Computing and Big Data Analysis (HPC), 891: 84–97, 2019.

[44] H. Silva, T. Basso, R. Moraes, D. Elia, S. Fior, "A re-identification risk-based anonymization framework for data analytics platforms," in Proc. European Dependable Computing Conference (EDCC): 101-106, 2018.

[45] K. Abouelmehdi, A. Beni-Hessane, H. Khaloufi, "Big healthcare data: Preserving security and privacy," J. Big Data, 5: 1-18, 2018.

[46] J. Domingo-Ferrer, J. Soria-Comas, "Anonymization in the Time of Big Data," International Conference on Privacy in Statistical Databases (PSD), 9867: 57–68, 2016.

[47] P. Ghavami, Big Data Analytics Methods: Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing, second ed., De Gruyter, Berlin, 2020.

[48] M. Z. Zgurovsky, Y. P. Zaychenko, Big Data: Conceptual Analysis and Applications, first ed., Springer Nature, Switzerland, 2020.

[49] D. Kumar Mishra, X. She Yang, A. Unal, Data Science and Big Data Analytics: ACM-WIR 2018 (Lecture Notes on Data Engineering and Communications Technologies, 16), first ed., Springer, Singapore, 2019.

[50] Rexa.info at the University of Massachusetts Amherst {Datasets Adult}.

[51] M. Kiabod, M. N. Dehkordi, B. Barekatain, "TSRAM: A Time-Saving k-degree Anonymization Method in Social Network," Expert Syst. Appl., 125: 378-396, 2019.

[52] A. Otgonbayar, Z. Pervez, K. Dahal, S. Eager, "K-VARP: k-anonymity for varied data streams via partitioning," Inf. Sci., 467: 238-255, 2018.

[53] G. Kaur, S. Agrawal, "Differential privacy framework: impact of quasi-identifiers on anonymization," in Proc. 2nd International

Conference on Communication, Computing and Networking, 46: 35–42, 2018.

[54] J. Wang, Z. Cai, Y. Li, D. Yang, J. Li, "Protecting query privacy with differentially private k-anonymity in location-based services," Pers. Ubiquitous Comput., 22: 453–469, 2018.

[55] Ch. N. Yang, Sh. L. Peng, L. C. Jain, Security with Intelligent Computing and Big-data Services, first ed., Springer Switzerland, 2020.

[56] L. Oneto, N. Navarin, A. Sperduti, D. Anguita, Recent Advances in Big Data and Deep Learning. Springer, Genova, 2020.

[57] J. Andrew, J. Karthikeyan, Privacy-Preserving Big Data Publication: (K, L) Anonymity, Advances in Intelligent Systems and Computing (AISC), 67: 77–88, 2020.

[58] Rexa.info at the University of Massachusetts Amherst {Datasets Bank and Marketing}.

[59] T. Banirostam, H. Banirostam, M. M. Pedram, A. M. Rahamni, "A review of fraud detection algorithms for electronic payment card transactions," J. Adv. Comput. Eng. Technol., 7(3): 157-166, 2021.

[60] H. Banirostam, E. Shamsinezhad T. Banirostam, "Functional control of users by biometric behavior features in cloud computing," in Proc. International Conference on Intelligent Systems, Modelling and Simulation: 94-98, 2013.

[61] H. Banirostam, A. Hedayati, A. Khadem Zadeh, E. Shamsinezhad, "A trust based approach for increasing security in cloud computing infrastructure," in Proc. UKSim-International Conference on Computer Modeling and Simulation: 717-721, 2013.

[62] H. Banirostam, A. R. Hedayati, A. Khadem Zadeh, "Using virtualization technique to increase security and reduce energy consumption in cloud computing," Int. J. Res. Comput. Sci., 4(2): 25-30, 2014.

[63] E. Shamsinezhad, A. Shahbahrami, A. Hedayati, A. Khadem Zadeh, H. Banirostam, "Presentation methods for task migration in cloud computing by combination of Yu router and post-copy," Int. J. Comput. Sci. Issues (IJCSI), 10(4): 98-102, 2013.

[64] T. Banirostam, E. Shamsinejad, M. M. Pedram, A. M. Rahamni, "A review of anonymity algorithms in big data," J. Adv. Comput. Eng. Technol., 7(3): 187-196, 2021.

[65] Z. El Ouazzani, H. El Bakkali, "A new technique ensuring privacy in big data: k-anonymity without prior value of the threshold k," in Proc. 1th International Conference On Intelligent Computing in Data Sciences, 127: 52-59, 2018.

[66] A. Raj, R. G L D'Souza, "Big data anonymization in cloud using k-anonymity algorithm using map reduce framework," Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., 5(1): 50-56, 2019.

[67] P. Jain, M. Gyanchandani, N. Khare, " Improved k-anonymize and l-diverse approach for privacy preserving big data publishing using MPSEC dataset," Comput. Inf., 39(3): 537–567, 2020.

Biographies

Elham Shamsinejad is a Ph.D. student in the Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran. Her research interests include Machine learning, Deep learning, Big Data, Data Analytics and Python Programming.

• Email: e.shamsinejad.eng@iauctb.ac.ir
• ORCID: 0009-0001-4941-8921
• Web of Science Researcher ID: NA
• Scopus Author ID: NA
• Homepage: NA

Touraj Banirostam is an Assistant Professor in the Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran. His research interests include Cognitive Science Engineering, Artificial Intelligence, Learning, Self-Management Systems.

• Email: h.banirostam.eng@iauctb.ac.ir
• ORCID: 0000-0002-3477-9046
• Web of Science Researcher ID: NA
• Scopus Author ID: NA
• Homepage: https://ctb.iau.ir/faculty/t-banirostam-comp/en

Mir Mohsen Pedram received the B.Sc. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, 1990, and the M.Sc. and Ph.D. degrees in electrical engineering from Tarbiat Modares University, Tehran, Iran, in 1994 and 2003, respectively. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Kharazmi University. His main areas of research are intelligent systems, machine learning, data mining, and cognitive science.

• Email: pedram@khu.ac.ir
• ORCID: 0000-0002-0674-4428
• Web of Science Researcher ID: NA
• Scopus Author ID: NA
• Homepage: https://eng.khu.ac.ir/cv/318/en

Amir Masoud Rahmani is currently working as a Professor for Islamic Azad University, science and research branch, Tehran. He is the author/co-author of more than 220 publications in technical journals and conferences. His research interests are in the areas of distributed systems, wireless sensor networks, Internet of Things and evolutionary computing. Address: Amir Masoud Rahmani, Computer Engineering dept, Islamic Azad University, Science and Research branch, Hesarak, Ashrafi Esfahani, Poonak Square, Tehran, IRAN.

• Email: rahmani@srbiau.ac.ir
• ORCID: 0000-0001-8641-6119
• Web of Science Researcher ID: NA
• Scopus Author ID: NA
• Homepage: https://eng.khu.ac.ir/cv/318/en

How to cite this paper:
E. Shamsinejad, T. Banirostam, M. M. Pedram, A. M. Rahmani, "Presenting a model of data anonymization in big data in the context of in-memory processing framework," J. Electr. Comput. Eng. Innovations, 12(1): 79-98, 2024.

DOI: 10.22061/jecei.2023.9737.651

URL: https://jecei.sru.ac.ir/article_1927.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Comprehensive Review of Modern Computing Paradigms Architectures for Intelligent Agriculture

M. Farmani, S. Farnam, M. J. Khani, Z. Torabi, Z. Shirmohammadi *

Department of Computer Engineering, Shahid Rajaee Teacher Training university, Tehran, Iran.

Article Info

Article History:
Received 11 March 2023
Reviewed 15 June 2023
Revised 10 July 2023
Accepted 23 July 2023

Keywords:
Intelligent agriculture
Computing technology
Cloud computing
Fog computing
Edge computing

*Corresponding Author's Email Address:
shirmohammadi@sru.ac.ir

Abstract

Background and Objectives: With the increase in population in the world along with the decrease in natural resources, agricultural land, and the increase of unpredictable environmental conditions, causes concerns in the field of food supply, which is one of the serious concerns for all countries of the world. Therefore, the agricultural industry has moved towards smart agriculture. Smart agriculture uses the Internet of Things, which uses different types of sensors to collect data (such as temperature, humidity, light, etc.), a communication network to send and receive data, and information systems to manage and analyze data. Smart agriculture deals with a huge amount of data collected from farms, which has fundamental challenges for analysis using old systems such as lack of storage space, and processing delay. The Computational paradigm is a key solution to solve the problems of time delays, security, storage space management, and real-time analysis. Computing paradigms include cloud, fog, and edge computing, which by combining each of them in smart agriculture has caused a great transformation in this industry. The purpose of this article is to provide a comprehensive review of the architecture of computing paradigms in smart agriculture applications.
Methods: To achieve the goals of this article, the methodology is divided into two parts: article selection and review of the selected articles. The computational paradigms used in the selected articles are from 2019 to 2022. Each selected paper is then reviewed in detail in terms of categories of computing paradigms, architectures, key points, advantages, and challenges.
Results: Computational paradigms have significant advantages. Combining these paradigms in a complementary way covers many challenges. The architecture based on the combination of edge-fog-cloud computing is one of the best architectures combined with smart agriculture.
Conclusion: By combining computing paradigms and smart agriculture, the challenges based on traditional and old systems are overcome. Combining these paradigms complement each other's challenges.

Introduction

Agriculture plays an essential role in the global food supply chain, which is the basis of human survival. According to the prediction of the World Health Organization, the population will reach 10 billion people by 2050 [1]. Therefore, if these predictions happen, the production of agricultural products in the universe should increment by about 60% annually [2]. In the last two decades, with the expansion of the Internet, there have been unlimited changes for organizations and citizens around the universe [3]. The appearance of the Internet of Things, which is defined as a network of objects in

which instruments, sensors, software, machines, and people are used through the Internet to communicate, exchange information, and interact between the real and virtual universe [4]. Internet of Things (IoT) systems also use various technologies such as Wireless Sensor Networks (WSN), cloud computing, and artificial intelligence. The Internet of Things is used in different scopes such as intelligent homes, healthcare, traffic, intelligent cities, and agriculture. Therefore, farmers, scientists, and agriculture industries have turned to intelligent agriculture, which uses methods and technologies at distinct levels and scales in the production of agricultural products.

In last years, intelligent agriculture has become very popular. Intelligent agriculture uses new technologies to maximize the use of wellsprings and minimize environmental impacts. Wireless Sensor Networks are one of these technologies that help farmers in the accumulation of information [3]. In intelligent agriculture, different sensors are used to accumulate data such as temperature, humidity, light, pressure, etc., which uses a correlation network to send and receive information, and finally, by analyzing the obtained information, it increments productivity and Minimizes waste is done at the right time and place [5]. Fig. 1, shows an example of IoT applications in intelligent agriculture. Existing sensors provide the complete status of agriculture products with accurate measurements. Based on the provided values, actuators manage agriculture processes related to beasts, crops, irrigation, etc. This can lead to predicting crop harvest, increasing production, reducing operating costs, remote monitoring, and accurate evaluation of farms [5].

Intelligent agriculture deals with plenty of heterogeneous information wellsprings.

Heterogeneous instruments and sensors accumulate agriculture information such as temperature, humidity, soil conditions, etc. Next, different actuators such as ventilation instruments, water supply systems, etc., adopt operations based on the information. With the development of science and technology, new methods and technologies have been presented in agriculture. An emerging trend is the use of the Internet of Things and Modern computing paradigms such as cloud, edge, and fog [5]. Computer-based agriculture systems have challenges such as processing acceleration, infrequent storage space, reliability, scalability, etc., which are unable to meet today's needs [6]. To solve these problems, services based on cloud services are used. Information captured by sensors is analyzed and processed using cloud services to make better decisions. Fog computing can also lead to declined network load and computing and storage in cloud servers. The purpose of this paper is to revise the existing investigation in the scope of computing technology architecture based on edge, cloud, and fog in intelligent agriculture. The structure of this work is as follows: In the second section, an overview of the concepts of computing technologies such as edge computing, fog computing, and cloud computing is presented. In the third section, the introduction of intelligent agriculture and its distinct scopes and the protocols used for correlation in it are examined.

In the fourth section, we review new investigations and describe the architecture of each of them. In the fifth section, the benefits and challenges of computing technology in intelligent agriculture are reviewed. Finally, the conclusion is given in the sixth section.

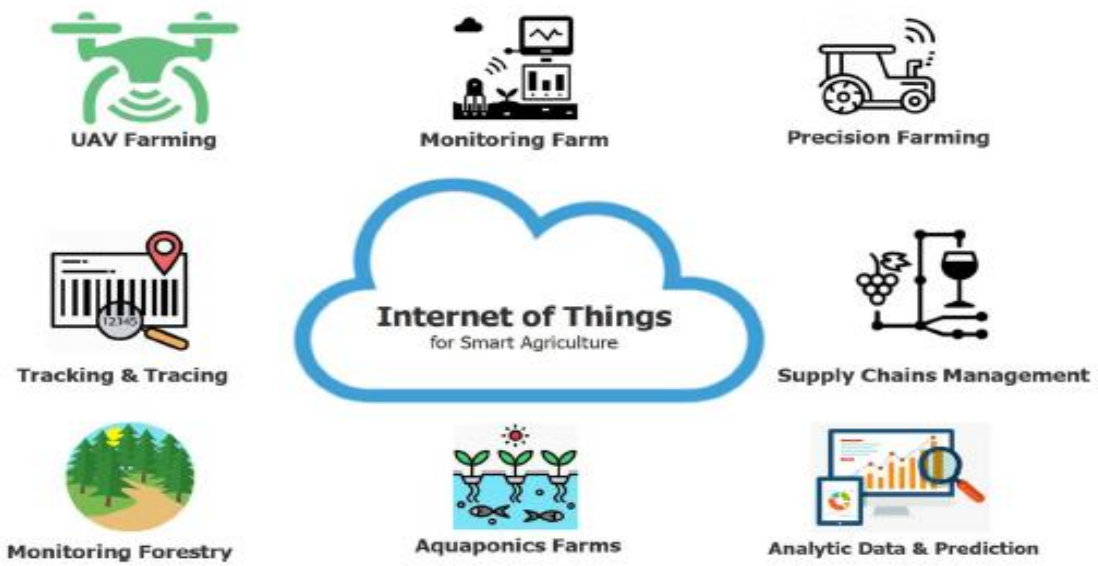


Fig. 1: Internet of Things applications in intelligent tillage [4].

An Overview of Modern Cloud Computing Technologies

A. Cloud Computing

In last years, cloud computing is becoming a principal technology in the scope of data technology. The phrase cloud computing was first used by Google and Amazon in 2006 [7]. Cloud computing is a model for easy access to computing wellsprings such as networks, servers, storage wellsprings, programs, and services through the Internet to provide access quickly and with minimum directorship requirements. Cloud computing includes five basic features, three service layers, and four distinct deployment models. The five necessary characteristics of the cloud consist of On-demand self-service, broad network access, wellspring pooling, Rapid elasticity and scalability, and measured service. A cloud must contain all five characteristics. The three service layers refer to the services provided by cloud providers and the user chooses them based on their needs. These three layers include infrastructure as a service (IAAS), software as a service (SAAS), and platform as a service (PAAS). Four cloud deployment models are also divided into Public Cloud, Private Cloud, Community Cloud, and Hybrid Cloud [8]. Clouds are formed by centralized servers that are also called information centers. Its advantages include fast deployment, cheap maintenance cost, and availability of stored information anywhere in the universe, simultaneous information analysis, and high computing power. But when dealing with big information, it has challenges such as time delay, internet bandwidth, real-time analysis, information directorship, and security [9].

B. Fog Computing

The new method of fog computing was proposed by Flavio Bonomi in Cisco in 2012 [10].

In fog computing, processing, and storage instruments are located close to every other and provide computing and storage services between end instruments and cloud computing information centers. Therefore, the basic idea is that the user's processing operations are performed in the nearby cloud and then sent to the cloud information centers. Fog computing is used for applications that require real-time processing with very infrequent time delays. Fog computing is distributed on a large scale and deployed where edge instruments perform processing [11], [12].

Fog nodes are an accumulation of nodes that receive information from IOT instruments in real-time. These nodes process the received information in less than a few milliseconds and periodically send analytical information to the central cloud. Every cloud node is equipped with internal computing wellsprings, information storage, networking, and information directorship and acts as a bridge between the central cloud layer and the edge layer [13]. The advantages of the fog node include infrequent delay, real-time interaction, mobility support, improved security, efficiency, and maintaining network bandwidth.

C. Edge Computing

Edge computing is a distributed architecture that enables computing at the edge of the network. In other words, computing is closer to the wellspring of information generation, that is, information is captured at the place where computing instruments perform analysis on them. Finally, the information extracted from the analysis is sent to the central cloud using the Internet. One of the principal advantages is that critical processes are monitored in real-time and operations are executed accordingly.

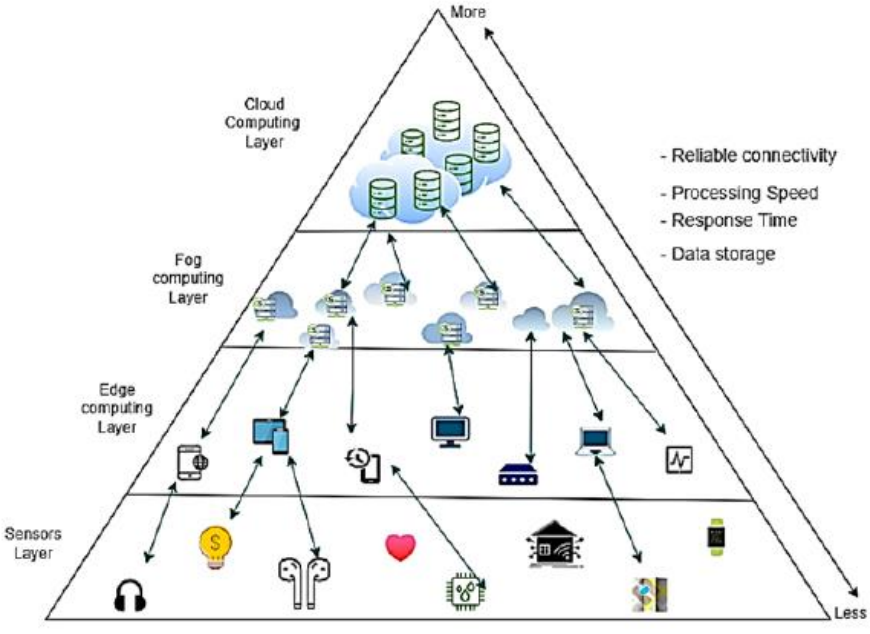


Fig. 2: Viewpoint of hybrid solutions for cloud, fog, and edge computing [14].

The advantages of using edge computing include very infrequent latency, high processing speed, and wide bandwidth.

These three computing concepts can be used in a complementary way. Fig. 2, shows an example of hybrid solutions for calculations. For instance, we can transfer easy processes to the edge nodes and assign a higher level of burdensome processing to the fog nodes and assign the processing of huge information to the cloud. The composition of these three clouds, fog, and edge computing creates maximum productivity in operations and applications [14].

An Overview of Intelligent Agriculture

A. Intelligent Agriculture

Intelligent agriculture is based on the knowledge and principles of agriculture, which uses a new method of planting, harvesting, and maintaining products and livestock products based on intelligent technologies. The purpose of intelligentization is to create a system to help decision-making in farms, which has advantages such as increasing production, saving energy, reducing manpower and increasing the efficiency of wellsprings, and improving the quantity and quality of products, etc. In intelligent agriculture, different technologies such as sensors, robots, and software are used, which provide farmers with positioning and information analysis. Farmers can monitor the ongoing activities anywhere in the universe and make the best decisions about farm activities with the information obtained from the intelligent system. Combining modern technologies with agriculture can lead to increment efficiency, and sustainability, which can have a significant impact on the universe's agricultural economy.

In last years, IoT applications have been used in different scopes of agriculture. These programs are mainly classified as agriculture programs such as crop directorship, greenhouse directorship, soil and water directorship, and livestock agriculture programs such as beast monitoring and livestock disease monitoring. In 2021, [5] categorized the main areas of intelligent agriculture into six categories, including crop directorship, beast directorship, irrigation directorship, soil directorship, climate directorship, and greenhouse directorship, which we will briefly describe below.

• Agriculture directorship: This directorship includes all the activities that are used to improve the growth and development of crop performance. Using sensors, drones, and intelligent robots, products can be managed and in case of pests and diseases, lack of water, the risk of harmful insects, etc., farmers can be informed about these harmful risks. To make the necessary decisions for better productivity of products [5].

• Beast directorship: Beast directorship includes all activities such as health, breastfeeding, breeding, which are done by farmers to raise farm beasts. To control the situation by installing distinct sensors on the beasts, their performance is checked at any moment based on specified factors, and if there is a threat, a warning will be sent to the farmers [15].

• Irrigation directorship: Irrigation directorship includes all activities that are used for proper planning and optimal use of water wellsprings. For water directorship, additional irrigation costs can be avoided by installing multiple sensors in appropriate places and leading to saving water wellsprings [5].

• Soil directorship: Soil monitoring is one of the environmental issues that have a great impact on crop production. For soil directorship, soil patterns such as humidity, temperature, deal of fertilizer, etc. are monitored. Suitable soil increments crop production [15].

• Weather directorship: Continuous weather monitoring is one of the most principal functions in agriculture. In this regard, tools are used to obtain weather parameters such as temperature, humidity, wind direction, air pressure, etc. The obtained information is used to improve agriculture productivity [15].

• Greenhouse directorship: Plants are grown in a greenhouse under controlled conditions. This directorship includes all the activities that are carried out to accurately control the environmental conditions suitable for growing plants [2].

B. Correlation Protocols in Intelligent Agriculture

In intelligent agriculture, many wireless correlation protocols are used based on the situation and available features. The instruments in the intelligent agriculture system can interact, switch data, make decisions for monitoring, control agriculture conditions, and amend performance and efficiency by using protocols. These protocols can be divided into short-range and long-range based on the correlation range. Short-range protocols include Bluetooth, ZigBee, and Radio Frequency Identification, and long-range protocols include Long Range (LoRa), SigFox, and Narrowband IoT (NB-IoT).

• Bluetooth: It is one of the wireless protocol technologies known by the IEEE 802.15.1 standard. This technology is infrequent cost and infrequent consumption and is used for transmission in a short range of 8 to 10 meters. Bluetooth acts in the 2.4 GHz frequency band [16]. Information transfer speed in distinct versions is from 1 to 24 Mbps.

• ZigBee: It is an IEEE 802.15.4 standard for wireless correlation designed for sensors and controls. This technology has a long battery life and is used for transmission up to a distance of 1 km [17].

• Radio Frequency Identification (RFID): This technology is suitable for long-range correlation. In this technology, every object has a unique identifier

separately and tracks and records the location of each of them [18].

• Universe Wide Interoperability for Microwave Access (WiMAX): It is an IEEE 802.16 standard that can cover a range of 50 km radius. The information transfer speed in this technology can increment up to 1 Gbit/s [19].

• Wireless Fidelity (WiFi): It is one of the local wireless network standards that use the Internet to transmit information wirelessly. This technology is known by IEEE 802.11 standard. Currently, it is one of the most widely used wireless technology in different instruments such as mobile phones, laptops, and tablets. Its coverage range is from 20 to 100 meters. The information transfer speed of this technology can be up to 700 Mbps. Of course, in distinct Wi-Fi standards, the coverage area and speed are distinct [19].

• SigFox: It is one of the wireless cellular networks that is suitable for long-distance correlation. This is an inexpensive network technology with infrequent power consumption and limited information rate and operates in a frequency band between 860 and 920 MHz. Its coverage range is from 10 to 50 km and the information transfer speed in this technology is up to 600 bits per second.

• Long Range (LoRa): It is a long-range wireless correlation technology that has very infrequent energy consumption and operates in an unlicensed band. Its coverage range is about 20 km and the information transfer speed in this technology is up to 100 kbps.

• Narrowband IoT (NB-IoT): It is an infrequent-power and infrequent-consumption long-range correlation technology introduced by the 3GPP standardization organization. This technology acts in infrequent bandwidth and covers a range of up to 35 km. The speed of information transfer in this technology is up to 250 Kb/s [20].

• Cellular correlation: It is one of the principal correlation technologies in the applications the Internet of Things that can transmit multimedia. These technologies include 3G, LTE, 4G, and 5G versions, which have wide cellular coverage, high throughput, and infrequent latency. This technology operates in the frequency band of 865 MHz and 2.4 GHz. The information transfer speed in this technology is distinct in distinct versions, for example, in 4G, the information transfer speed varies from 100 Mbps to 1 Gbps [21].

Types of Computing Technology Architecture

In the last halls, the vast investigation has been carried out in the scope of intelligent agriculture with edge, fog, and cloud computing technologies. These investigations have distinct combinations of computing technologies that can be divided into 4 categories including intelligent agriculture and cloud computing, intelligent agriculture and a combination of edge and cloud computing, intelligent agriculture and a combination of fog and cloud computing, intelligent agriculture and a combination of edge computing- Fog - the cloud split. In the following, we will examine every of these categories in the last investigations, which can be seen in Table 1, a summary of the reviewed investigations.

A. Architecture Based on Cloud Computing

In [22], an infrequent-cost intelligent system for monitoring environmental parameters using drones and cloud computing technology is presented. In this system, it periodically gathers information using a soil moisture sensor; these sensors forward data to the gateway. Then, using a drone equipped with long-range network technology, the obtained information is sent to the cloud. Finally, in the cloud, the operation of accumulation and storing remote user information, information processing and displaying the outcomes to the user, analyzing the information obtained from the sensor, and making decisions are done. Fig. 3, shows the architecture of the suggested system. This intelligent system based on cloud computing can help farmers analyze the information on environmental conditions in large farms, leading to increment crops, better directorship, and time-saving.

In [23], an Internet of Things system based on long-range network and cloud computing is suggested for intelligent farms. According to Fig. 4, the suggested system consists of four portions namely sensor nodes, control equipment, clouds server, and a web application platform. Sensor nodes are distributed across the scopes and accumulate information about the state of the scope and send it to cloud servers for information storage and directorship. Cloud server communicates with sensors and warehouse control network using long-range network. In the warehouse control network, a Programmable Logic Controller (PLC) is used to control the process and drive instruments. The cloud server receives sensor information using long-range network gateways and stores it in databases. Farmers can remotely monitor the system using any intelligent instrument such as mobile phones and laptops through the monitoring program on the cloud server using a Web browser, control as well as review captured information. This design has outcompeted in high scalability, increment number of sensors for monitoring, increment efficiency, and remote directorship. In [24], a Wireless Sensor Network system based on cloud computing is suggested for monitoring farm beasts. This system stores the locations of livestock movements in real-time. The monitoring system includes three parts: Wireless Sensor Network, cloud platform, and user interface, as shown in Fig. 5. The Wireless Sensor Network segment includes sensors for beast monitoring. Every sensor obtains location information at a specific time using the Global Positioning System (GPS).

Table 1: Review of investigation acts based on computing technologies

Reference	Year	Suggested method	Architecture			Protocol
			Edge layer	Fog layer	Cloud layer	
[27]	2022	Monitoring system based on the Internet of Things and edge and cloud computing for intelligent farm	Sensors, Edge Gate	-	Cloud server	WSN, Wifi
[23]	2022	Internet of things system based on long-range network and cloud computing in intelligent tillage	-	-	Cloud server	LoRa
[24]	2021	Wireless Sensor Network system based on cloud computing for beast monitoring	-	-	Cloud services	WSN, 3G
[22]	2021	Intelligent system for monitoring environmental parameters in intelligent tillage	-	-	Cloud server	LoRa
[5]	2021	Cloud-fog-edge computing model for intelligent tillage	Sensors, Actuators, Tractors	Fog node	Cloud server	LoRa, ZigBee, SigFox, Bluetooth
[30]	2020	Monitoring system for fire prediction combining cloud and fog computing for environment and tillage	-	Fog node	Cloud server	ZigBee
[32]	2020	Intelligent knowledge system architecture based on edge-fog-cloud computing	Edge node, Edge gate	Fog node, Fog gate	Cloud services	-
[28]	2020	Architecture based on fog computing and long-range network technology in intelligent farm	Sensors, Actuators	Fog node	Cloud	LoRa
[29]	2020	A deep learning method to fog nodes in cloud-based intelligent tillage	-	Fog node	Cloud server	-
[26]	2020	Information accumulation method using edge computing in intelligent greenhouse	Sensors, Edge servers	-	Central Cloud	ZigBee, wifi
[25]	2019	Home edge computing architecture	Sensors, Home edge computing, Multiple Access	-	Central Cloud	LoRa
[31]	2019	Advanced system for remote tillage monitoring using long-range network and edge-fog-cloud computing combinations	End instrument, Edge gates	Fog gates, Repeaters	Cloud server	LoRa

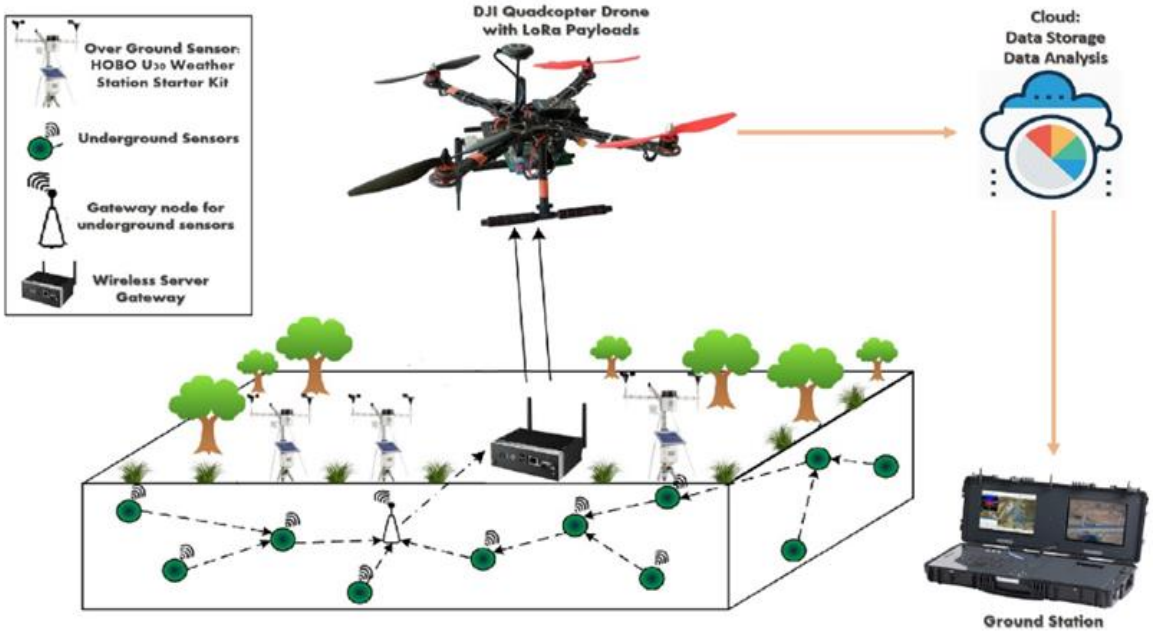


Fig. 3: Infrequent-cost intelligent system architecture for farm monitoring using UAV and cloud computing [22].

Then the information generated by the sensors is captured and sent to the second part for information storage, directorship, and processing. In the second part, processing is done on the information in cloud services, and uploads the obtained information on a web page. Cloud computing platforms provide computing power, information storage, and applications. In the third section, the user interface provides the current system status and processed information. Farm managers can check the estate of every beast through website pages and analyze the comportment of each of them using this system.

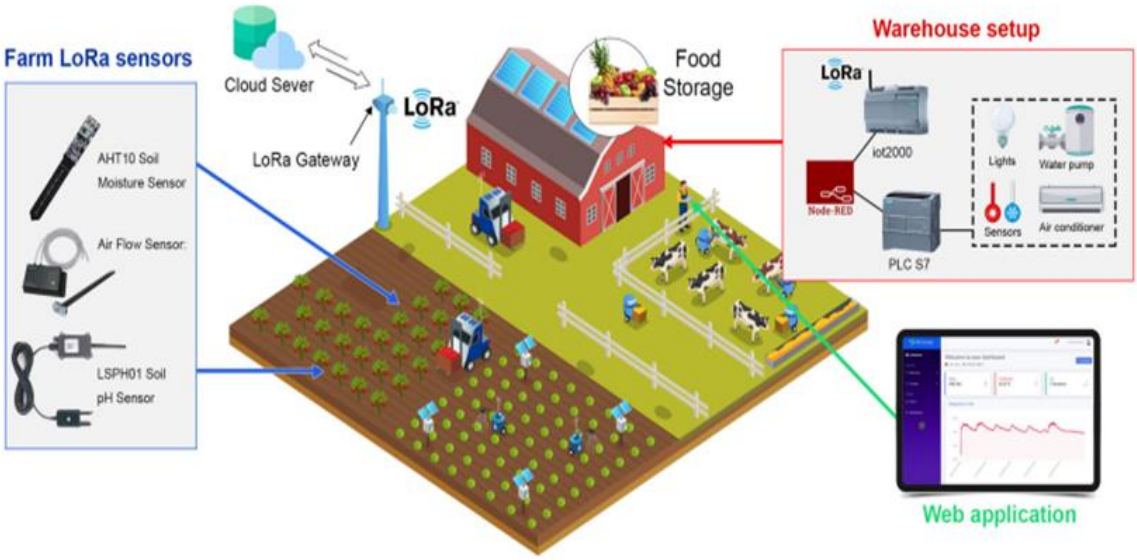


Fig. 4: Internet of Things system based on long-range network and cloud computing for intelligent farm [23].

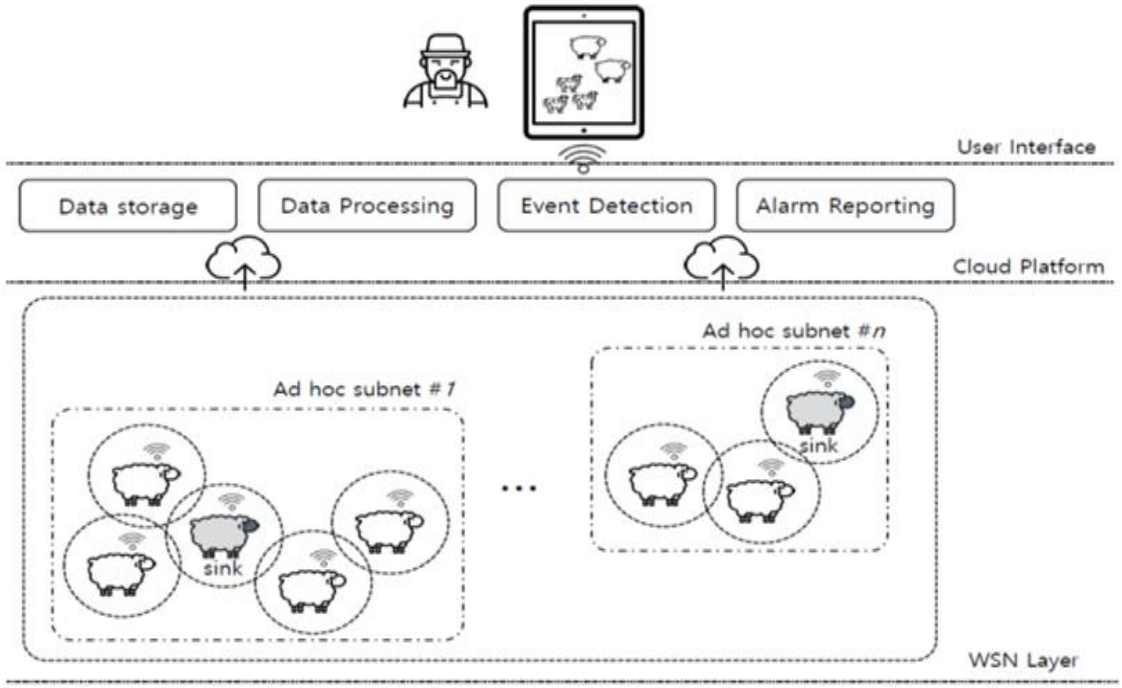


Fig. 5: Architecture based on cloud computing for monitoring beasts [24].

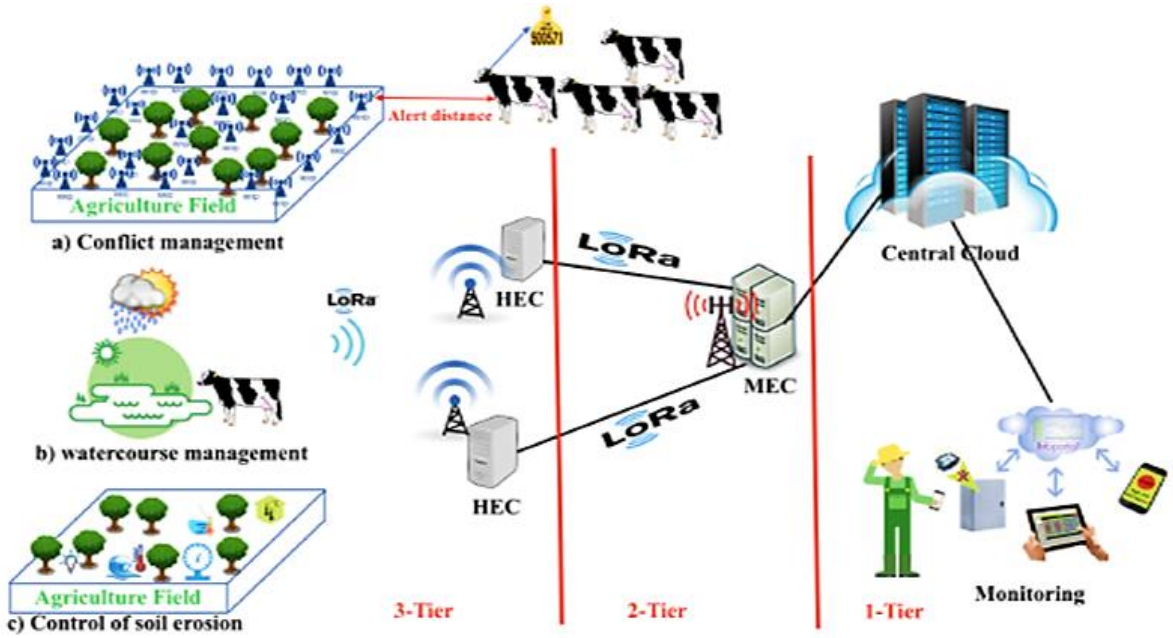


Fig. 6: Three-layer architecture of home edge computing in intelligent tillage [25].

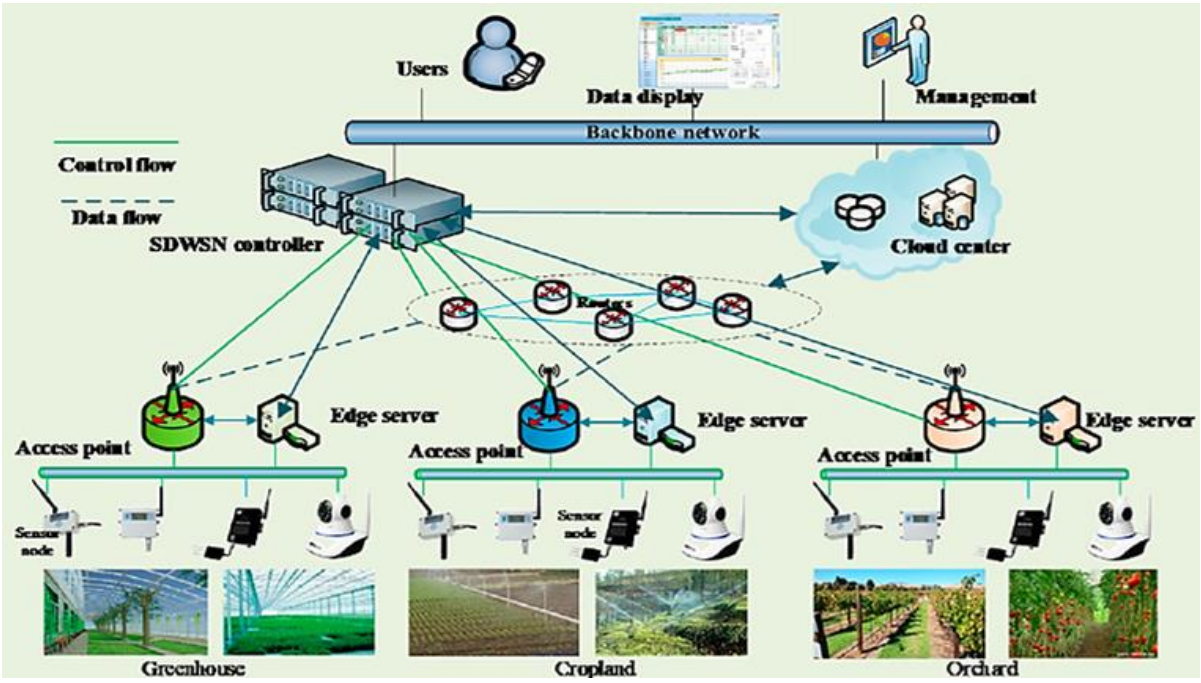


Fig. 7: The architecture of information accumulation method with the help of edge computing in intelligent tillage [26].

B. Architecture Based on a Combination of Edge and Cloud Computing

In [25], a Home Edge Computing (HEC) architecture for intelligent and tolerable agriculture and breeding is suggested. This architecture consists of three layers and is based on home edge computing. Fig. 5, shows the architecture related to home edge computing in intelligent agriculture.

This architecture includes three levels of cloud which are local cloud or home server, edge cloud, and central cloud. This architecture is suggested to solve the latency quandary in Multi-Access Edge Computing (MEC) for certain kinds of applications that require too high availability of wellsprings and must be processed with very infrequent delays. According to Fig. 6, in the third level, sensors and instruments related to intelligent agriculture are located in this layer, which is connected to

a local information center called home edge computing. Home edge computing performs local information processing and acts as a gateway to higher levels. The duty of this gateway is that if further wellsprings are needed, it transfers the traffic to higher levels, i.e., multi-access edge computing at the second level and central cloud at the first level. If the distance between the two sites of multi-access edge computing and home edge computing is far, they can be connected to every other through a point-to-point radio correlation using a long-range network protocol. This architecture leads to the reduction of delay in the network.

In [26], an information accumulation method with the help of edge computing for principal events and reducing information redundancy in intelligent agriculture is suggested. This method consists of four parts including Wireless Sensor Network, Software-Defined Wireless Sensor Network (SDWSN) layer, edge computing layer, and application layer. Fig. 7, shows the architecture of this method. The Wireless Sensor Network layer includes different sensors in the scope of agriculture and access points. The function of access points is to create effective links between sensors and edge servers in the cloud. Software-Defined Wireless Sensor Network layer has been used to increment the flexibility of the system for information accumulation. By using cloud computing, information processing, and storage capacity are provided for the application layer. Cloud-based applications are also divided into categories of information visualization, user demand analysis, and system directorship. The overall method as shown in Fig. 7, is that the cloud server obtains the characteristics of principal events by analyzing and processing the information. Then, the edge server determines the information received from the sensors and the feature value of principal events and performs optimization on the information based on the features of the principal event, and the corresponding information is sent to the Software-Defined Wireless Sensor Network. In the next step, the sensor nodes receive the information related to the measured information and the correlation parameters considering the principal events from the Software-Defined Wireless Sensor Network. In the next step, the Software-Defined Wireless Sensor Network sends control streams including information accumulation commands and principal parameters to the access points. Next, this information is sent to cloud centers for directorship. Finally, by checking the information on the clouds, the outcomes are sent to the application layer. This method leads to a reduction in correlation time and infrequent delay in the information accumulation system.

In [27], a hybrid monitoring system based on the Internet of Things and edge and cloud computing is suggested for an intelligent farm. This system has 3 main layers including an accumulation layer, a decision layer, and an application layer as illustrated in Fig. 8.

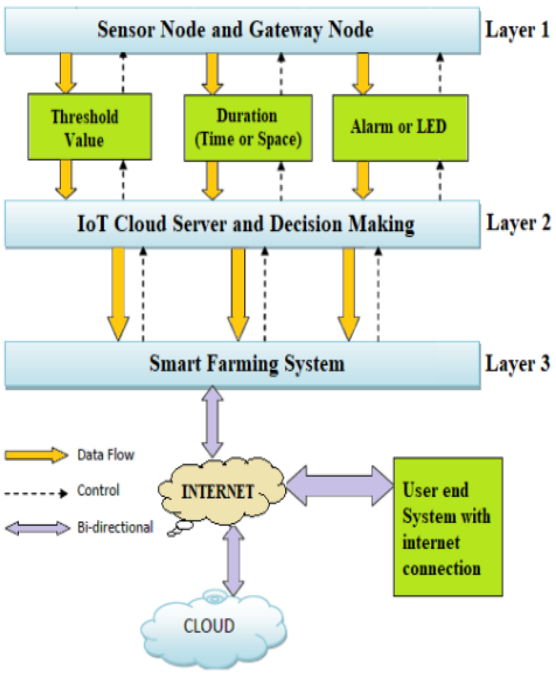


Fig. 8: Architecture of hybrid monitoring system based on edge and cloud computing [27].

In the first layer, sensors are deployed in distinct monitoring locations. All sensing data from distinct scopes in the scope is captured and stored using wireless network gateway nodes at the edge. Then the sensor information is sent to the second layer for decision-making and storage in the cloud server. In this layer, values are defined as threshold limits for all types of sensors. At a certain time, this layer examines all the obtained data using threshold values and makes appropriate decisions. Finally, in the application layer, related decisions for different scopes are communicated to the end user through SMS, email, and website. The user can access the website anywhere in the universe and make an effective decision to monitor the farm. This system leads to increasing production, increasing productivity, reducing the cost of producing products, and increasing the speed of processes.

C. Architecture Based on a Combination of Fog and Cloud Computing

In [28], an architecture based on fog nodes and long-range network technology is proposed to optimize the number of sensors deployed in an intelligent farm. Fig. 9, shows the architecture based on fog nodes and long-range network technology. According to Fig. 9, the sensors and actuators in the smart farm are connected to fog nodes.

The role of the fog node in this architecture is to create a bridge between the sensors and the network and is located as a local server near the data source. It also manages data collected from sensors. Then, in the next step, the fog nodes transmit only the important information to the cloud via the Internet. In this proposed architecture, the fog nodes process and store the generated data of the sensors locally and prevent all the information grown from the sensors from moving to the cloud. This action reduces network delay and information processing can be done in real time. The use of fog computing in this architecture has improved real-time processing, reduced latency, and saved bandwidth.

In [29], a deep learning method for fog nodes in intelligent agriculture based on cloud computing is suggested. The target of this technique has been to minimize the response latency and further processing of deep learning tasks for intelligent agriculture applications. This method consists of several layers and every layer deal with the input information from the prior layer to extract features and produce an outcome to deliver to the subsequent layer.

The lowest layer handles the incoming pure information and the highest layer provides the processed data at the output. Every layer declines the information volume for delivery to the subsequent layer. Fig. 10, shows the general picture of the deep learning method in intelligent agriculture. A cloud server assigns several layers of the deep learning model to fog nodes and maintains the remaining layers. Every instrument has information related to the fog node close to it, and at the request of applications, it transfers the captured information to the corresponding fog node. By receiving raw information from instruments, a fog node performs the defined layers of the deep learning model corresponding to an application. After the deep learning layers process is completed in the fog node, the outcomes are sent to the central cloud server, which performs the remaining deep learning layers for the outcome. As further layers are assigned to the fog nodes, the deal of information sent to the cloud through the network decreases, and eventually the network congestion and computing load on the cloud are declined.

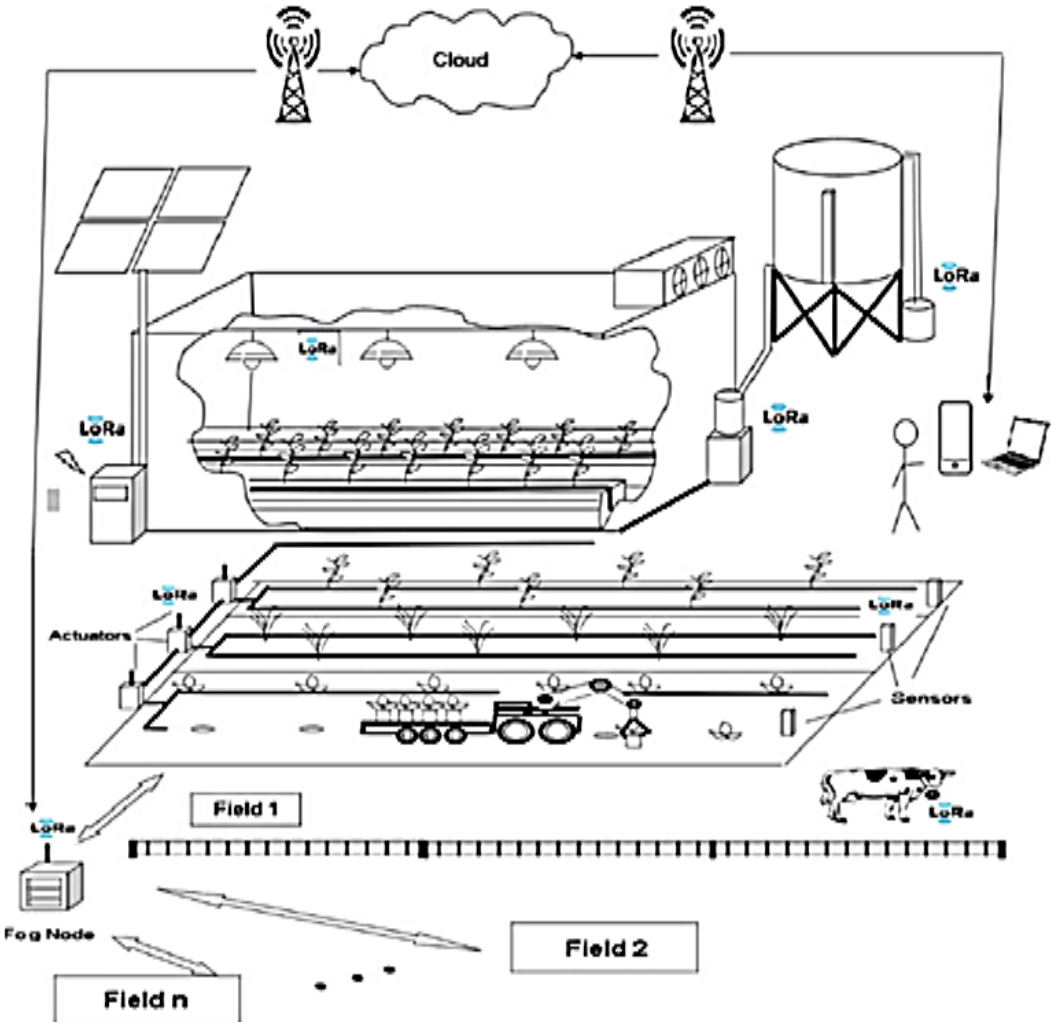


Fig. 9: Architecture based on fog node and long-range network technology in the intelligent farm [28].

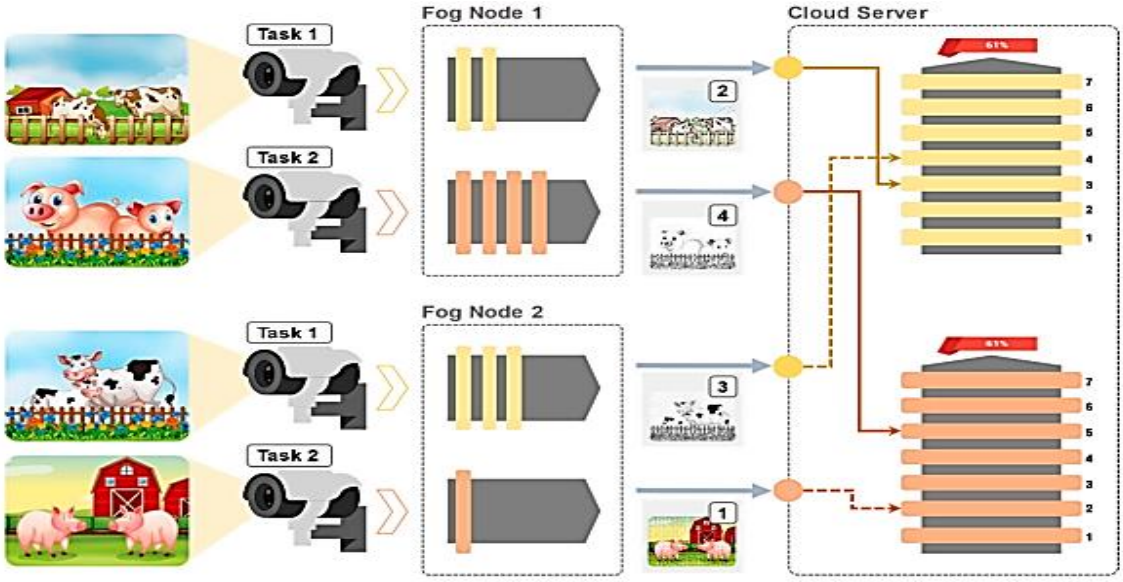


Fig. 10: The overall pcture of the deep learning method to fog nodes in intelligent tillage [29].

In [30], an Internet of Things system based on a Wireless Sensor Network with a combination of cloud and fog computing is suggested for fire prediction and early detection in the environment and agriculture sectors. This suggested system has 3 layers as illustrated in Fig. 11. The first layer consists of a cloud infrastructure formed by information centers that are dynamically allocated facilities and wellsprings based on user's requests. The services of this layer include information storage and processing. At the second layer is cloud computing, which can work as small information centers that are more inexpensive and more accessible, extending service delivery to the edges of the network. This act declines the computational load, frees up wellsprings, prevents network traffic, and increments system capacity. In the third layer, fire directorship events are addressed using agriculture and environmental monitoring programs. For example, there are instruments called sensory nodes that monitor environmental factors such as temperature and humidity. Next, using a sink node accumulations information from sensors and transmits them to higher layers for more calculations. Finally, end users such as farmers, fire services, etc., through graphical user interfaces, intelligent instruments, or mobile phones, perform the necessary actions in emergencies. Using fog nodes can decline latency, increment throughput, and control energy consumption.

D. Architecture Based on a Combination of Edge-Fog-Cloud Computing

In [31], an advanced long-range grid-based system using a mixture of edge, fog, and cloud computing is suggested for remote agriculture, which is also applicable for use in remote areas that are developing.

The architecture of this system includes five layers, including the sensor layer, edge layer, fog layer, cloud layer, and final layer, as shown in Fig. 12. This system should be used for situations that have infrequent power transmission and long-range with limited information rate. The sensor layer includes several groups of sensor nodes and actuator nodes. These nodes are deployed in distinct areas of the farm according to their applications. The sensor nodes send the captured information to the edge gateways at the edge layer, while the actuator nodes receive commands from the edge layer for control. In the second level, the edge layer consists of edge gateways and is responsible for receiving information from sensor nodes. Then, after processing the information in this layer, the processed and compressed information is sent to the fog gateways in the fog layer. The edge layer has many benefits including fast notification, channel categorization, and security. The edge layer and the fog are connected to every other through broadband network technology, which can transmit information at an infrequent speed of 10 to 20 kilometers. On the third level, the fog layer, includes two parts, repeaters, and fog gates. In this layer, the information sent from the edge gateways is received and sent to the fog gateways. The section of repeaters is used to maintain the correlation link so that packets are not lost to transmit information over long distances. The fog gateways section is also used to share sensor information. The advantages of this layer include advanced services such as distributed information storage, information fusion, information processing, and security. At the fourth level, it consists of the cloud layer, which includes cloud servers and their services, such as global information storage, major information analysis,

and information processing with complex algorithms. The final layer, also called the end-user layer, includes mobile phone and web browser applications that are used to access real-time information and provide input commands for remote farm control.

In [32], an intelligent knowledge system based on a combination of edge, fog, and cloud computing is suggested for farmers to use in intelligent farms. This system helps farmers in making decisions to increment the production and profits of crops. Farmers can connect with this system using mobile applications and web applications and receive the required information from expert experts. Fig. 13, shows the architecture of the intelligent knowledge system. This system includes five layers, including the agriculture surroundings layer, edge computing layer, fog computing layer, cloud computing layer, and intelligent user interface layer. In the agriculture environment layer, sensors are used to monitor environmental parameters. This layer accumulates information from sensors and sends it to the node in the edge computing layer. The edge computing layer, also includes an edge node that accumulates the information sent by the sensor node and processes them. Using this layer in the network has the advantages of reducing delay and reducing traffic in the cloud network. The fog computing layer, it includes fog nodes, which are in the appearance of servers and storage instruments. This layer has three operations of information accumulation, information display for analysis, and knowledge generation for farmers.

The task of this layer is to classify and filter information so that the deal of information transmissive to the cloud is declined and information processing is close to knowledge production. Then the information is sent to the cloud computing layer. In this layer, different information of the sent information and agriculture land are stored, which are performed on them and stored in it. Finally, the intelligent interface layer, is an interface between farmers and the intelligent agriculture system. Users can act with the intelligent farming system using web pages or intelligent phones to get comprehensive data about the agricultural land under observation. The suggested system by using the combined computing of edge, fog, and cloud has brought advantages such as increasing efficiency, reducing delay, reducing cost, high scalability, and increasing speed and security in intelligent agriculture.

In [5], a computing model based on cloud fog edge is suggested for intelligent agriculture. In this model, the cloud layer is used to store information, analyze information in huge volumes, and upload algorithm and information analysis tools to the fog node, storing backup information for future analysis. On the second level are fog layers that are installed on local farms. This layer is responsible for real-time information analysis, decision making, and information reasoning. After analyzing and processing the information, it is sent to the cloud layer for more analysis and backup. The third level, which is the edge layer, consists of end instruments, tractors, sensors, and actuators. In Fig. 14, the architecture of this computational model is shown.

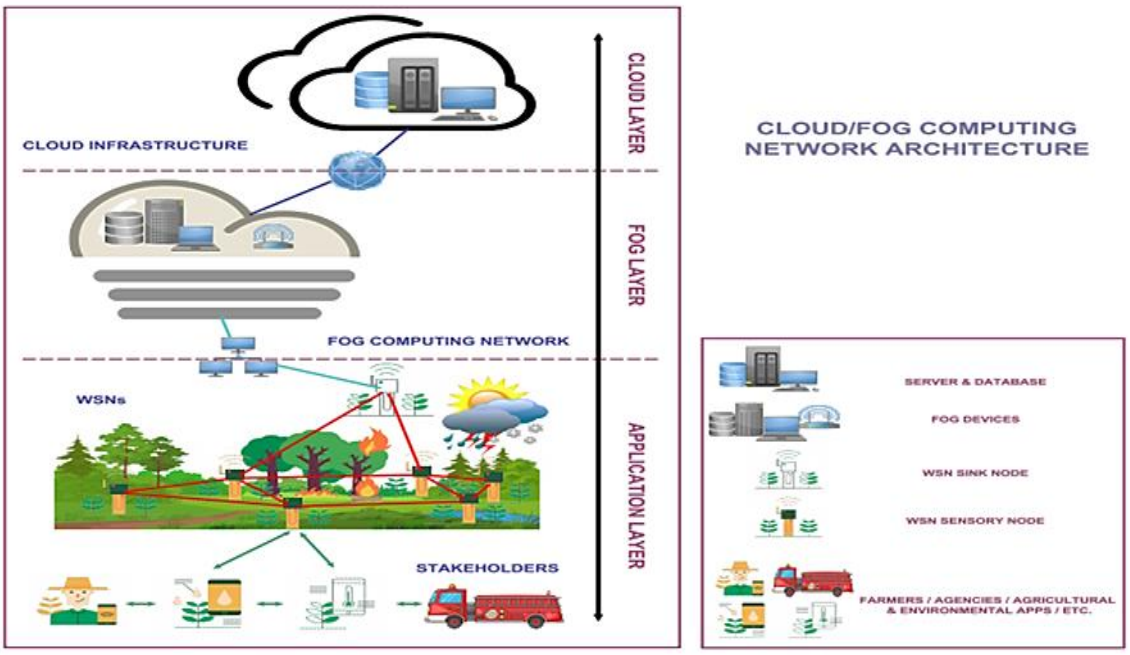


Fig. 11: Architecture of the Internet of Computing Objects with the combination of cloud and fog for environmental and tillage monitoring [30].

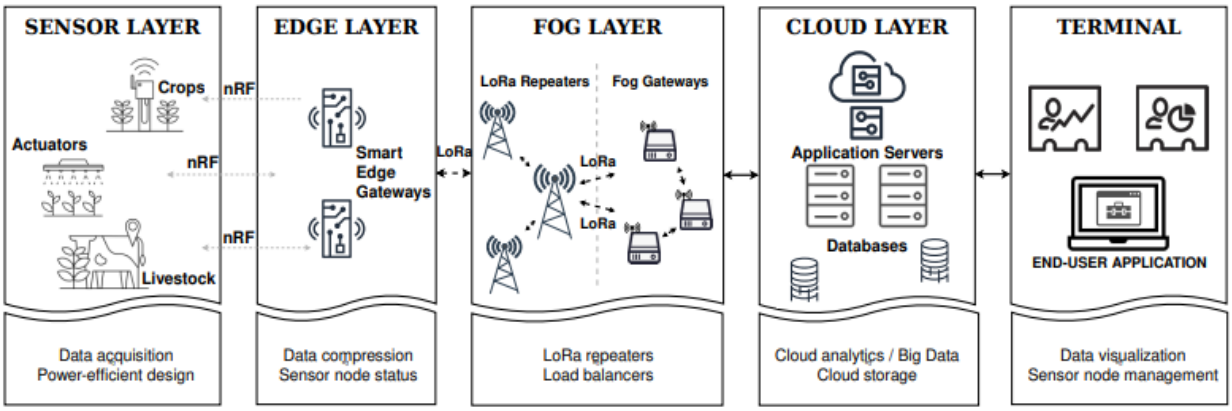


Fig. 12: Advanced system architecture based on a long-range network with a mixture of edge, fog, and cloud computing [31].

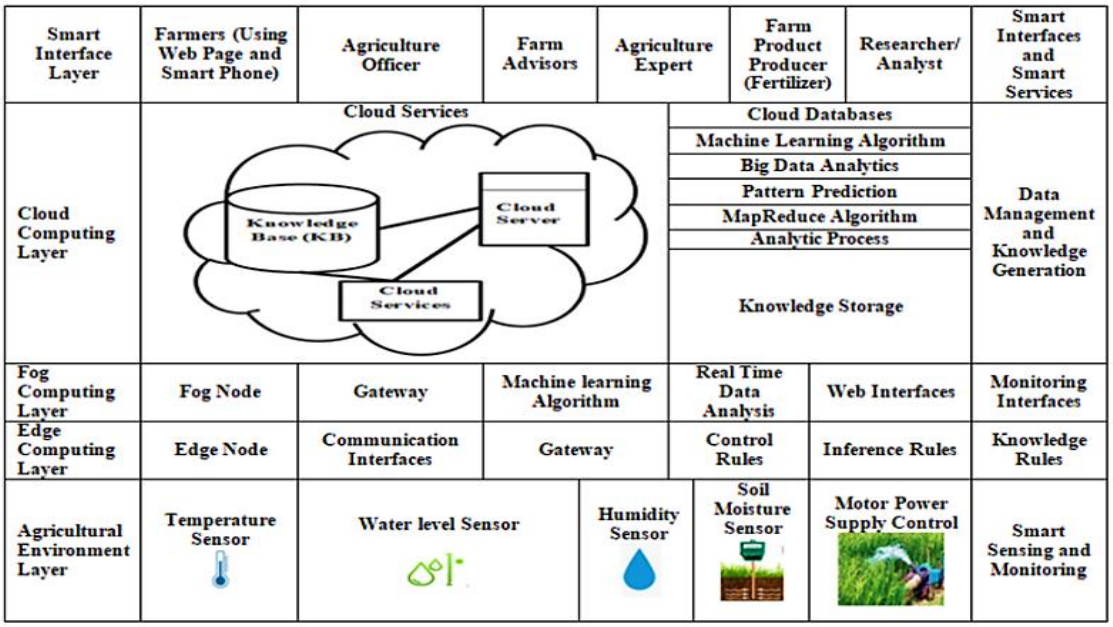


Fig. 13: Intelligent knowledge system architecture based on edge, fog, and cloud combinations [32].

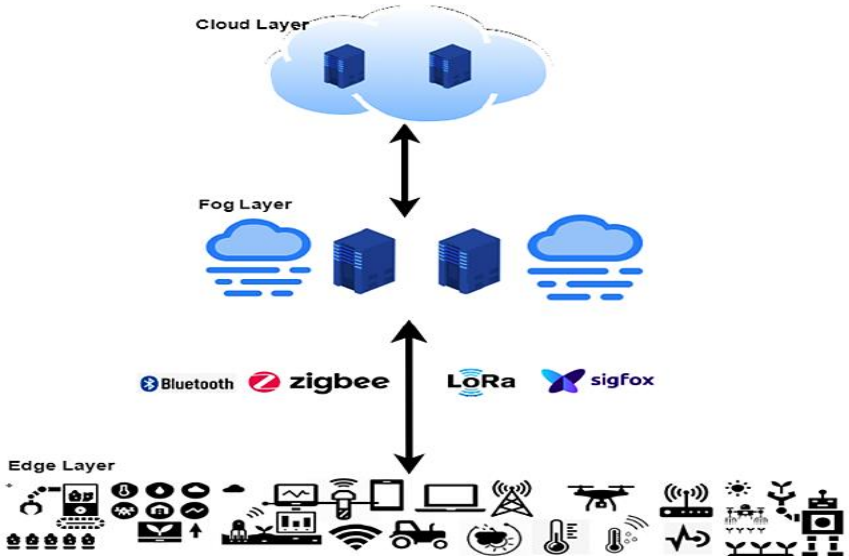


Fig. 14: Three-layer architecture based on cloud-fog-edge calculations in intelligent tillage [5].

Advantages and Challenges of Computing Technology in Intelligent Agriculture

Edge, fog, and cloud-based technologies have significant advantages in the agriculture sector. Some of its advantages are 1- the improvement of information directorship which is done by the service provider and categorizes the information in an organized manner, 2- users can access the information at any moment and position, 3- the user is free from repair and Infrastructure maintenance is safe because service providers are responsible for technical issues, 4-building and improving the supply chain of agriculture products, mentioned [33]. In contrast, these technologies also have significant challenges. Among these challenges are 1-security and privacy, which may cause great economic losses to farmers and industries if the information stored in the cloud servers is transferred outside. 2- intelligent farms should not go to the cloud for information analysis depend because it is not suitable for real-time information processing, 3- Intelligent farms need fast support and real-time information processing to be able to accumulate more information from farms, which is not possible just by connecting to the cloud, 4- Poor internet is one of the main challenges are in intelligent farms, because it can cause information loss, delay in information processing, decrease the speed of information loading, he pointed out [5]. Using fog and edge computing, problems caused by real-time processing, reducing delay and increasing bandwidth, increasing information security, and local information processing can be solved.

Conclusion

Agriculture is one of the principal parts of the universe's economy and human life. Intelligent agriculture using the Internet of Things tries to decline the problems of traditional agriculture and increment the production of agriculture products and accumulate information about the current situation for farmers. With the emergence of new technologies such as computing technologies and their combination with intelligent agriculture, the challenges of intelligent agriculture based on old systems will be overcome. In this article, the architecture of edge, fog, and cloud computing technologies, advantages, and challenges of this technology were investigated. These computing technologies can complement every other and cover many challenges. Smart agriculture can solve the challenges of traditional agriculture, but it faces many challenges, including the energy of sensors and the challenges of deploying sensors and data security. In this article, the challenges of smart agriculture were examined and the work that can be done in the future to solve these challenges was explained.

Author Contributions

Mojtaba Farmani, Saman Farnam and Zahra Shirmohammadi contributed to the idea, review, writing and editing paper. Mohammad Javad Khani writing and editing paper. Zeinab Torabi and Zahra Shirmohammadi edited/reviewed the paper.

Acknowledgment

The authors would like to thank the anonymous reviewers and the editors of JECEI for their valuable comments and suggestions for improving quality of the paper. This work was supported by Shahid Rajaee Teacher Training University under grant number 4894 and 4898.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

IAAS	Infrastructure As A Service
SAAS	Software As A Service
PAAS	Platform As A Service
SDWSN	Software-Defined Wireless Sensor Network
PLC	Programmable Logic Controller
HEC	Home Edge Computing

References

[1] U. Nations, growing at a Slower Pace, World Population is Expected to Reach 9.7 billion in 2050 and Could Peak at Nearly 11 billion around 2100.

[2] N. Alexandratos, J. Bruinsma, World Agriculture towards 2030/2050: The 2012 Revision. 2012. (accessed on 1 September 2021).

[3] M. S. Farooq, S. Riaz, A. Abid, T. Umer, Y. Bin Zikria, "Role of IoT technology in agriculture: a systematic literature review," Electronics, 9(2): 319, 2020.

[4] V. K. Quy, N. V. Hau, D. V. Anh, N. M. Quy, N. T. Ban, S. Lanza, A. Muzirafuti, "IoT-Enabled smart agriculture: architecture, applications, and challenges," Appl. Sci., 12(7): 3396, 2022.

[5] Y. Kalyani, R. Collier, "A systematic survey on the role of cloud, fog, and edge computing combination in smart agriculture," Sensors, 21(17): 5922, 2021.

[6] S. Singh, I. Chana, R. Buyya, "Agri-Info: cloud based autonomic system for delivering agriculture as a service," Internet Things, 9: 100131, 2020.

[7] M. De Donno, K. Tange, N. Dragoni, "Foundations and evolution of modern computing paradigms: Cloud, IoT, edge, and fog," IEEE Access, 7: 150936-150948, 2019.

[8] The NIST Definition of Cloud Computing, NIST Special Publication 800-145, 2011.

[9] E. Symeonaki, K. G. Arvanitis, D. D. Piromalis, "Review on the trends and challenges of cloud computing technology in climate-Smart agriculture," in Proc. HAICTA: 66–78, 2017.

[10] F. Bonomi, R. Milito, J. Zhu, S. Addepalli, "Fog computing and its role in the internet of things," in Proc. the first edition of the MCC

Workshop on Mobile Cloud Computing: 13–16, 2012.

[11] R. K, Naha, S. Garg, A. Chan. “Fog computing architecture: Survey and challenges,” arXiv preprint arXiv:1811.09047, 2018.

[12] R. Deng, R. Lu, C. Lai, T. H. Luan, H. Liang, "Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption," IEEE Internet Things J., 3(6): 1171-1181, 2016.

[13] A. Yousefpour, C. Fung, T. Nguyen, K. Kadiyala, F. Jalali, A. Niakanlahiji, J. Kong, J. P. Jue, “All one needs to know about fog computing and related edge computing paradigms: A complete survey,” J. Syst. Archit., 98: 289–330, 2019.

[14] S. Dhifaoui, C. Houaidia, L. A. Saidane, “Cloud-Fog-Edge computing in smart agriculture in the Era of drones: a systematic survey,” in Proc. 2022 IEEE 11th IFIP International Conference on Performance Evaluation and Modeling in Wireless and Wired Networks (PEMWN): 1-6, 2022.

[15] M. S. Farooq, S. Riaz, A. Abid, K. Abid, M. A. Naeem, “A survey on the role of IoT in agriculture for the implementation of smart farming,” IEEE Access, 7: 156237-156271, 2019.

[16] L. Ruiz-Garcia, L. Lunadei, P. Barreiro, I. Robla, “A review of wireless sensor technologies and applications in agriculture and food industry: State of the art and current trends,” Sensors, 9: 4728–4750, 2009.

[17] M. Bacco, A. Berton, A. Gotta, L. Caviglione, “IEEE 802.15.4 air-ground UAV communications in smart farming scenarios,” IEEE Commun. Lett., 22: 1910–1913, 2018.

[18] X. Wang, J. Zhang, Z. Yu, S. Mao, S. C. G. Periaswamy, J. Patton, “On remote temperature sensing using commercial UHF RFID tags,” IEEE Internet Things J., 6: 10715–10727, 2019.

[19] S. Popli, R.K. Jha, S. Jain, “A survey on energy efficient Narrowband Internet of Things (NBIoT): Architecture, application & challenges,” IEEE Access, 7: 16739–16776, 2019.

[20] T. Ojha, S. Misra, N. S. Raghuwanshi, “Wireless sensor networks for agriculture: The state-of-the-art in practice and future challenges,” Comput. Electron. Agric., 118: 66-84, 2015.

[21] O. Ali, M. K. Ishak, M. K. L. Bhatti, I. Khan, K. I. Kim, “A comprehensive review of internet of things: Technology stack, middlewares, and fog/edge computing interface,” Sensors, 22: 995, 2022.

[22] F. A. Almalki, B. O. Soufiene, S. H. Alsamhi, H. Sakli, “A low-cost platform for environmental smart farming monitoring system based on IoT and UAVs,” Sustainability, 13(11): 5908, 2021.

[23] M. Saban, O. Aghzout, A. Rosado-Muñoz, "Deployment of a LoRa-based network and web monitoring application for a smart farm," in Proc. 2022 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT): 424-427, 2022.

[24] J. K. Park, E. Y. Park “Monitoring method of movement of grazing cows using cloud-based system,” ECTI Trans. Comput. Inf. Technol., 15(1): 24-33, 2021.

[25] C. S. M. Babou, B. O. Sane, I. Diane, I. Niang, “Home edge computing architecture for smart and sustainable agriculture and breeding,” in Proc. the 2nd International Conference on Networking, Information Systems & Security: 1-7, 2019.

[26] X. Li, Z. Ma, J. Zheng, Y. Liu, L. Zhu, N. Zhou, “An effective edge-assisted data collection approach for critical events in the SDWSN-based agricultural internet of things,” Electronics, 9(6): 907, 2020.

[27] M. A. Uddin, U. Kumar Dey, M. Akter, "Proposing a cloud and edge computing based decision supportive consolidated farming system by sensing various effective parameters using IoT," in Proc. 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS): 1-6, 2022.

[28] M. Baghrous, A. Ezzouhairi, N. Benamar, “Smart farming system based on fog computing and lora technology,” Embedded Systems and Artificial Intelligence, 1076: 217-225, 2020.

[29] K. Lee, B. N. Silva, K. Han, “Deep learning entrusted to fog nodes (DLEFN) based smart agriculture,” Appl. Sci., 10(4): 1544, 2020.

[30] A. Tsipis, A. Papamichail, I. Angelis, G. Koufoudakis, G. Tsoumanis, K. Oikonomou, “An alertness-adjustable cloud/fog IoT solution for timely environmental monitoring based on wildfire risk forecasting,” Energies, 13(14): 3693, 2020.

[31] T. N. Gia, L. Qingqing, J. P. Queralta, Z. Zou, H. Tenhunen, T. Westerlund, “Edge AI in smart farming IoT: CNNs at the edge and fog computing with LoRa,” in Proc. 2019 IEEE AFRICON: 1-6, 2019.

[32] U. Sakthi, J. D. Rose, “Smart agricultural knowledge discovery system using IoT technology and fog computing,” in Proc. Third International Conference on Smart Systems and Inventive Technology (ICSSIT): 48-53, 2020.

[33] E. Symeonaki, K. G. Arvanitis, D. D. Piromalis, “Review on the trends and challenges of cloud computing technology in climate-smart agriculture,” in Proc. HAICTA: 66-78, 2017.

Biographies

Mojtaba Farmani received the B. Sc. degree in Computer Engineering from Shahid Rajaee Teacher Training University in 2020. He is a M.Sc. graduate student in Computer software from Shahid Rajaee Teacher Training University. He is currently a researcher in the Wireless sensor network, Wireless body sensor network, patient diagnosis, energy consumption and data prediction.

- Email: mojtabafarmani@sru.ac.ir
- ORCID: 0009-0009-6950-5860
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Saman Farnam received the B. Sc. degree in Computer Engineering from Shahid Shamsipour University. He is a M.Sc. graduate student in Computer software from Shahid Rajaee Teacher Training University. He is currently a researcher in the Wireless sensor network, Wireless body sensor network, energy consumption and power management.

- Email: samanfrnam@gmail.com
- ORCID: 0009-0006-1403-7228
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mohammad Javad Khani received B.Sc. in Computer Engineering (Software) from Shahid Rajaee Teacher Training University, Tehran, Iran, and M.Sc. in Computer Engineering (Software) from Shahid Rajaee Teacher Training University, Tehran, Iran. He is a lecturer professor at Qom Technical and Vocational University. His main research interests are Sampling, Compressing, Wireless Body Area Networks, Wireless Sensor Networks, Energy efficiency, Smart Agriculture.

- Email: m.j.khani1375@gmail.com
- ORCID: 0000-0002-3643-7824
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Zeinab Torabi received her Ph.D. degree in Computer Architecture from Shahid Beheshti University, Tehran, Iran, in 2016. She is currently an Assistant Professor in Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, and Tehran, Iran. Her research interests include computer arithmetic, residue number system, and algorithms.

- Email: z.torabi@sru.ac.ir
- ORCID: 0000-0002-2526-688X
- Web of Science Researcher ID: ABG-9144-2022
- Scopus Author ID: 56958405600
- Homepage: https://www.sru.ac.ir/en/school-of-computer/zeinab-torabi/

Zahra Shirmohammadi received M.Sc. and Ph.D. degrees in Computer Engineering from Sharif University of Technology in 2011 and 2017 respectively. Her current research interests include dependability of System-onChip (SoC) and Network-on-Chip (NoC) design and high-performance computer architecture.

- Email: shirmohammadi@sru.ac.ir
- ORCID: 0000-0003-2607-4940
- Web of Science Researcher ID: ABD-8084-2020
- Scopus Author ID: 56039488300
- Homepage: https://www.sru.ac.ir/en/school-of-computer/zahra-shirmohammadi/

How to cite this paper:
M. Farmani, S. Farnam, M. J. Khani, Z. Torabi, Z. Shirmohammadi, "Comprehensive review of modern computing paradigms architectures for intelligent agriculture," J. Electr. Comput. Eng. Innovations, 12(1): 99-114, 2024.

DOI: 10.22061/jecei.2023.9682.648

URL: https://jecei.sru.ac.ir/article_1928.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Design, Analysis, and Implementation of a New Online Object Tracking Method Based on Sketch Kernel Correlation Filter (SHKCF)

M. Yousefzadeh, A. Golmakani *, G. Sarbishaei

Department of Electrical Engineering, Sadjad University of Technology, Mashhad, Iran.

Article Info

Article History:
Received 08 July 2023
Reviewed 17 August 2023
Revised 11 September 2023
Accepted 11 October 2023

Keywords:
Artificial intelligence
Video analysis
Object tracking
SHKCF
KCF and online tracker

*Corresponding Author's Email Address:
golmakani@sadjad.ac.ir

Abstract

Background and Objectives: To design an efficient tracker in a crowded environment based on artificial intelligence and image processing, there are several challenges such as the occlusion, fast motion, in-plane rotation, variations in target illumination and Other challenges of online tracking are the time complexity of the algorithm, increasing memory space, and tracker dependence on the target model. In this paper, for the first time, sketch matrix theory in ridge regression for video sequences has been proposed.
Methods: A new tracking object method based on the element-wise matrix with an online training method is proposed including the kernel correlation Filter (KCF), circular, and sketch matrix. The proposed algorithm is not only the free model but also increases the robustness of the tracker related to the scale variation, occlusion, fast motion, and reduces KCF drift.
Results: The simulation results demonstrate that the proposed sketch kernel correlation filter (SHKCF) can increase the computational speed of the algorithm and reduces both the time complexity and the memory space. Finally, the proposed tracker is implemented and experimentally evaluated based on video sequences of OTB50, OTB100 and VOT2016 benchmarks.
Conclusion: The experimental results show that the SHKCF method obtains not only OPE partial evaluation of Out of view, Occlusion and Motion Blur in object accuracy but also achieved the partial evaluation of Illumination Variation, Out of Plane Rotation, Scale Variation, Out of View, Occlusion, In of Plane Rotation, Background Clutter, Fast Motion and Deformation in object overlap which are the first rank compared to the state-the-art works. The result of accuracy, robustness and time complexity are obtained 0.929, 0.93 and 35.4, respectively.

Introduction

One of the most important aspects in the machine vision and pattern recognition systems is object tracking which has different applications such as video surveillance with CCTV, medical and military video analysis, human-computer communication, and robotic smart vehicle [1]-[6], [12]. The main task of tracking is that the object location is automatically estimated and scaled in the video sequences [1]. In the last decade, mathematical theories and modelling techniques in object tracking have been developed. These techniques are based on the dynamic and static learning theory, particle filters, discriminate correlation filters, pattern matching, deep neural networks, etc. [5].

Despite this impressive progress, the existence of an object tracking algorithm that can be adapted to different conditions is still a challenging problem in machine vision. Besides, the majority of most previous approaches have been designed and implemented in simple landscapes. However, the online tracking of real objects in

unpredictable landscapes is still a big challenge because of illumination variations, occlusion, scale variations, and background clutter [6]-[8]. Therefore, developing this new design of online tracker is still needed to open up practical tracking applications. It is worth mentioning that the challenges based on tracking benchmarks can help to assess the accuracy, precision, and robustness of the tracking algorithm instead of the challenges within the real-world environments and online video sequences [9], [11]. For examples, OTB50 and OTB100 are two well-known famous standard benchmarks [5], [12], [13].

Generally, modern trackers are divided into two main categories: the generative and discriminating trackers [9]. The generative tracker method is based on the random variable probability estimating explained as follows. The object is labelled on the first frame. Then, the estimation error between the initial and new samples is calculated. Finally, based on the results of the previous step, the best candidate is generated. This approach extracts two models: an appearance model for the object and another model for the background. The main goal of this tracker is to predict the object location using the maximum similarity of the test sample to the appearance model. The tracker calculates the density of test samples around the object location to increase prediction accuracy. Then the object is selected by a particle filter. The main problem of the generative tracker is its dependency on the labeled datasets. Also, small changes in the background cause object estimation from the test samples to be inaccurate and very erroneous. As a to these challenges, researchers are aiming to use new trackers, such as the average transmission tracker [15], ensemble object tracker [15], fragment-based tracker [16], and sparse representation tracker [17]. However, the discriminating tracker can detect the object within the background regarding an online classified problem [2]. Nowadays, discriminating trackers deploy machine learning and tracking algorithms based on the kernel correlation filter. The main reasons for using the correlation filter-based trackers are their high speed, stability, and accuracy since these trackers use both the object and background information [7], [9], [11]. Despite the advancements of discriminating trackers, the efficiency of the designed tracking algorithms must be evaluated and compared with different objects on the existing benchmarks and challenges (e.g., OTB100 and OTB50).

The robustness and efficiency of the tracker are two main issues in the field of tracking that are highly competing in the literature. As already mentioned, improving the mentioned issues should be evaluated according to standard challenges (e.g., OTB50, OTB100, VOT2019, UAV123, LaSOT and TrackingNet). In this article, we have used two of the most widely used benchmarks, ot50 and otb100, and all our results are based on these two standard datasets. Also, the speed of the designed algorithm for calculating the object location is another issue that has to be considered within the algorithm evaluation. Recently, there have been attempts to overcome the above problems, such as multiple learning tracker [18], ensemble tracker [19], SVM tracker [8], correlation filter-based tracker, to name a few. It should be noted that the high capability and performance of the trackers based on discriminating correlation filters (DCF), compared to the most up-to-date available tracking algorithms, have been proven in [9], [18], [19]. In fact, the main advantages of DCF trackers are the multidimensional appearance model, circulant matrix, and frequency domain calculations explained in the next section.

In the frequency domain, DCF learning comes with a huge learning cost, such as circular shift samples of the ground-truth object, because a circular shift introduces the unwanted boundary effects. This problem is partially mitigated by additional predefined spatial constraints on the filter coefficients. For example, Danelljan et al. [52] introduced spatial regularized differential correlation filters (SRDCF) to reduce boundary effects. It is expected from an object tracker to be spatially penalized by its distance from the object center. In order to generate true positive and negative samples of the model training, Galoogahi et al. [50] proposed Learning background-aware correlation filters (BACF) for multiplied the correlation filter directly to binary matrix. In order to these mentioned method Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking (STRCF) and Visual Tracking via Adaptive Spatially-Regularized Correlation Filters (ASRCF) are also employed in literature for reducing mentioned boundary effect with spatial temporal Regularized and adaptive Spatially-Regularized, respectively.

It should be mention that the appearance model of most DCF-based trackers is developed by a linear interpolation approach. However, these models cannot adapt to overall appearance variances, leading to filter degradation. In order to solve the problem of filter degradation, several approaches have been proposed in the literature, such as the training set management [31], [32] and the temporal constraints, where temporal regularization has been proven to be an effective method.

In order to solve most to challenging problem of unwanted boundary effect, we proposed sketch kernel correlation filters (SHKCF) for real time object tracking. We implemented and compared our approach with state-of-the-art trackers in the OTB 100 benchmarks. The results show that SHKCF performs better than the state-of-the-art trackers in terms of accuracy and computational speed.

The rest of the paper is organized as follows: Section II presents the related works on video sequence trackers. Section III introduces the KCF tracking algorithm. Then, the proposed method based on the kernel ridge regression (KRR), sketch kernel ridge regression (SKRR), and SHKCF video sequence tracking algorithm is analyzed and presented. Section V provides the test results and compares them with the SOTA works regarding the OTB100 and OTB50 datasets.

Related Works

As already mentioned, appearance modeling is one of the most important approaches for Conventional object tracking, which can be classified roughly into discriminative and generative methods [37]. Generative approaches identify an object by learning reference model with the most similar video sequence region, including sparse representation [55], template matching [39], subspace learning [54], the generative methods can provide more accurate performance in a small region and are robust against the object occlusion. But they are sensitive to the same distractions in the object' surrounding region.

During the past decade, trackers based on the well-known regression are investigated and provided well performance [12]. Particularly, the series of the correlation filters-based trackers, KCF [1], SAMF [38], DSST [56], are demonstrated to be the best tracker in accuracy on the challenging of OTB100 [26].

Kernel correlation filter (KCF) is one of the methods used in DCF based trackers that has been highly investigated and developed by researchers in recent years [8]. The KCF method uses a set of patch mappings (positive/negative) and a classification between the object and its surroundings to create a kernel model. Fast Fourier transform (FFT) and the inverse fast Fourier transform (IFFT) are also employed to increase computational speed It is worth noting that the calculation speed of the correlation filter in the frequency domain is higher than in the time domain. Therefore, all calculations are performed in the frequency domain, and then the IFFT is employed to return them to the time domain. Finally, the KCF algorithm returns the object location as the output [9]. The results in [2], [5], [20] show that the KCF method enhances the object tracking performance compared to the SOTA works in terms of speed and robustness regarding the standard benchmark platforms [8]. However, some KCF method errors against various challenges remain unsolved, such as smart training, speed, drifting, and occultation. To the best of our knowledge, no other KCF object tracker has employed the sketch method. Although the sketch is introduced in Reference [7], its combination with the video sequence tracker has not yet been reported. So, in this paper, we combine the sketch method with the classic KCF tracker and demonstrate its pros and cons. In this paper, we propose a new method for object tracking based on the kernel correlation filter (KCF) and compare the results with the SOTA algorithms in object tracking scenarios. Our proposed method is based on SHKCF principle, which is proposed for the first time in video sequence tracking. We implement and evaluate our method according to OTB50 and OTB100 benchmarks to validate the results.

The main reason for the development of these algorithms is that the correlation filter calculates the learning coefficients and minimizes energy consumption [21]. This filter is also employed to calculate the variance of the training sample responses [22]. It should be noted that the basic filter-based trackers were designed based on calculations in the time domain. However, Bolme et al. proposed a filter learning method in the frequency domain [23]. In these filters, convolutional operators in the time domain transform into summations and multiplications in the frequency domain leading to a decrease in computational time. Therefore, implementing filters in the frequency domain leads to a very high frame rate, and the temporal complexity of the calculations reduces significantly. Although FFT-based algorithms are very effective and have many applications in signal processing, FFT has many limitations in tracking applications. Recently, the most efficient FFT-based tracker has been proposed by Henriques using the kernel method [1]. Previous algorithms did not consider a clear relationship between nonlinear kernels and Fourier domain parameters. So, they had high computational complexity and had limitations in image processing. This fact motivates Henriques to propose a simple connection between the transfer video patches and the training algorithms. Since 2015, researchers have improved Henriques' tracker through combination with other methods such as circular structure kernel (CSK) tracker, color name (CN) tracker, spatially scaled discriminating tracker, and kernel correlation filter (KCF). The CSK tracker is designed based on light intensity features, but it is not robust to some challenges such as occlusion and deformation [9]. The KCF deployed the density of samples around the object location, minimum kernel squares, and the rotational shifting structure of the video patch for learning [15]. This tracker shows a more accurate and robust performance than the CSK tracker, because the intelligent samples in the correlation filter are trained by both the histogram of gradient directional (HOG) features and the circular shift matrix on the video sequence. In addition to the HOG features, KCF can be combined with the color name feature to promote multi-feature detections. The KCF has significant computational properties leading to online frame per second (FPS) rates. The KCF tracker, similar to the other trackers, can be trained using neural networks and deep learning

algorithms. Recently, [24] proposed a deep learning-based method for estimating the object's location. This method, known as the CCOT method, combines location information with neural network features and correlation filters. It is worth noting that the CCOT tracker wins the VOT2016 challenge. For a fair comparison, we have compared the results of the proposed method with the results of CCOT using OTB100 and OTB50 challenges datasets. We also compared the results with the SOTA works having the same parameters and processing standard platform. Moreover, the KCF tracker can be combined with other mathematical matrices such as a multidimensional matrix with color scale features [25], [26] in a way that the calculations are performed using kernel functions and circulant matrix structure. To develop KCF trackers in the free model, multiple KCF trackers are proposed in [27]. Also, the online classifier Fern tracker is proposed in [28] to solve some other challenges (e.g., occlusion, out of view). To increase accuracy, some methods use the object and its vicinity pixels.

In other words, the KCF tracker intercepts the object features using circular shifts and takes samples in the vicinities of the desired object.

The Traditional KCF Tracker Algorithm

The pattern in the first frame of the video sequence (X) is assumed as the input circulant matrix. This one-dimensional circulant matrix is obtained by the original data set received from continuous frames: $P_x = [x_n . x_1 . \dots . x_{n-1}]^T$. Although, this circulant matrix can be extended to its two-dimensional form which considers all circular shifts: $\{P_x^i | i = 0 . \dots . n-1\}$. There are two possible shifts in each direction for this matrix. The matrix generated by the possible circular shifts is called the circulant matrix or data matrix. Therefore, the goal of KCF learning is to train the H filter as follows. Considering the minimum regression error, the KCF classifier is trained according to (1).

$$Arg \min_H \sum_i^n (f(H; P^i x) - Y_i)^2 + \lambda \|H\|_2^2 \quad (1)$$

$$f(H; P^i x) = H^T \Phi(X) \quad (2)$$

where f, H and $\Phi(X)$ are the mapping function, KCF filter and the mapping of the X pattern in the Fourier domain, respectively. In this equation, two patterns are used, the learning pattern (X) and the regression pattern (Y). This method is known as the minimum output sum of squared error (MOSSE) calculation.

The MOSSE calculation is obtained from the maximum values of the Gaussian function with the purpose of the minimal change in the circular shift. It should be noted that the digitalization of the Gaussian function, often termed P^i is achieved in the KCF method based on the degree of circular shifts. The digital matrix P^i is a matrix of zeros and ones indicating the incorrect and correct data, respectively. The detecting probability of an object is proportional to the number of correct data. Thus, for a detected object, the sum of correct data, is higher than a predefined threshold. The threshold is set by the designer and avoids falling into the local minimum trap. Equation (1) can be rewritten as (3):

$$\mathcal{L}(H) = \text{Arg} \min_H \|H^T \Phi_H - Y\|_2^2 + \lambda \|H\|_2^2 \quad (3)$$

where Φ_H , λ≥0 and Y are the mapping of all circular shifts of the X pattern in the Fourier domain, the regularization parameter and the results of the regression object pattern, respectively. Equation (3) shows the cost function depends on the partial mapping function Φ (.) explained as follow.

If the partial mapping Φ is linear, then $\Phi(X) = X$ where $X = [X_1\ X_2\ \dots]^T$. The KCF filter function (H) can be written by $\boldsymbol{H} = (X^T X + \lambda I)^{-1} X^T Y$, where $\boldsymbol{I}$ is the identity matrix. The components of the H filter can be obtained from the X pattern using a circulant matrix. In other words, the H filter performs the diagonal matrix calculations using the discrete Fourier transform matrix (DFT). Therefore, the H filter in the Fourier domain with Hadamard Product (element-wise multiplication) is expressed as:

$$\hat{H}^* = \frac{\hat{X}^* \odot \hat{Y}}{\hat{X}^* \odot \hat{X} + \lambda} \quad (4)$$

where $\hat{X}$ and $\hat{X}^*$ are the fast Fourier transform of X and the complex conjugate of X, respectively. It is worth mentioning when several patterns are employed in the training steps, the H filter is combined with all the patterns as [11]:

$$\hat{H}^* = \frac{\sum_{j=1}^m \hat{X}^* \odot \hat{Y}}{\sum_{j=1}^m \hat{X}^* \odot \hat{X} + \lambda} \quad (5)$$

Due to the linearity of partial mapping Φ in (2), it is not possible to calculate the multiple features of the objects. If the partial mapping Φ is nonlinear, then the X pattern has various properties such as HOG. Due to the nonlinearity of the mapping, (2) will not have a suitable response. Thus, to solve the nonlinear mapping, the problem is converted to the ridge regression method and the new H is calculated using kernel filter analysis as $H = \Phi^T \alpha$, where $\alpha = (K + \lambda I)^{-1} Y$ and $K = \Phi_X \Phi_X^T$.

The Gaussian kernel function creates multiple features of the X pattern. In [16], it is proved that if the kernel matrix has a constant permutation, the kernel matrix K is circular. Regarding the circularity of the K matrix and DFT of the diagonal property in (2), the coefficient α (frequency domain) can be extracted as [7].

$$\hat{\alpha}^* = \frac{\hat{Y}}{\hat{K}^{XX'} + \lambda} \quad (6)$$

For the well-known Gaussian kernel $\widehat{K}(\mathrm{x},\mathrm{x}') = \exp\left(-\frac{1}{\sigma^2}(\|\mathrm{x}-\mathrm{x}'\|^2)\right)$, Next

$$\widehat{K}^{XX'} = \exp\left\{-\frac{1}{\sigma^2}\left(\|X\|^2 + \|X'\|^2 - 2F^{-1}\left(\sum_d \hat{X}_d^* \odot \hat{X}_d'\right)\right)\right\} \quad (7)$$

where $\widehat{K}^{XX'}$, $\hat{X}$ and $\hat{Y}$ are the Gaussian kernel, the object of FFT filter and response, respectively.

A. Kernel Ridge Regression with Circulant Matrix

The idea of the proposed method is to minimize the difference between the output and the object's real location. To create a circulant matrix, we need to store the first column of the input matrix. Since a circulant matrix depends on its first column, a matrix formed by the circular shift method requires less memory [29]. The circulant matrix is shown as (8):

$$C = \begin{bmatrix} C_1 & C_m & C_{m-1} & \cdots & C_2 \\ C_2 & C_1 & C_m & \ddots & C_3 \\ C_3 & C_2 & C_1 & \ddots & C_4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_m & C_{m-1} & C_{m-2} & \cdots & C_1 \end{bmatrix} \quad (8)$$

The matrix (8) is expressed in closed form (9):

$$C_{[m]} = cir\left[c_j : j \in \{1.2.\ldots.m\}\right] \quad (9)$$

Besides, the matrix C with two inputs i and j is displayed in the simple closed-form solution of (10):

$$C_{ij} = c_{(i-j)} mod\ m \quad (10)$$

It should be noted that the advantage of a circulant matrix is not only the decreased memory but also calculations in the Fourier domain are faster compared to time-domain convolution. Discrete Fourier transform (DFT) of a circulant matrix is calculated by (11):

$$C = \frac{1}{m} G^* diag(Gc) G \quad (11)$$

where $C = [\mathrm{C}_1,\mathrm{C}_2,\ldots,\mathrm{C}_\mathrm{m}]^T$, $G = \left[e^{i\left(\frac{2\pi}{mkt}\right)}\right]_{k\cdot t=1}^{m}$ and G^* are circulant matrix transpose, discrete Fourier matrix and its conjugate, respectively. $diag(Gc)$ is a diagonal matrix whose diagonal elements are the elements of vector G. Moreover, the computational complexity is decreased from order m^2 to $mlog(m)$ by the proposed method for tracking [9], [18].

B. Complexity Analysis for Circulant Matrix

A circulant matrix $\mathrm{C} \in \mathbb{R}^{\mathrm{m}\times\mathrm{m}}$ [29] is a structured matrix, which is completely defined by its first column so that to reconstruct the entire matrix, need to store the first column, where m is the sketch dimension. The space complexity is $\mathcal{O}(m)$ instead of $\mathcal{O}(m^2)$. Therefore, the space complexity for solving $\mathcal{O}(nm)$ Furthermore, the circulant matrix can obtain a matrix-vector product ($\mathrm{C} * \mathrm{V}, \mathrm{V} \in \mathbb{R}^{\mathrm{m}}$)by the fast Fourier transform (FFT), whose time cost is $\mathcal{O}(m\ log(m))$ [44] the time required for the same operation with the unstructured Gaussian sketch. Therefore, the time complexity for solving the sketch matrix–kernel matrix product (SK) in our method is $\mathcal{O}(nm\ log(m))$. For details, see [11].

Most importantly, the effectiveness of a circulant matrix whose inputs in the first column are independent and identically distributed (i.i.d.) Gaussian inputs is almost the same as that of an unstructured matrix with i.i.d. Gaussian inputs [42]. Due to the advantages mentioned above, the circulant matrices have attracted extensive attention in some fields: approximation of the kernel matrices [43], [44], kernel selection [45], [46], approximation of the kernel function [41], binary embedding [47], ect. To the best of our knowledge, the circulant matrix based on the random sketch has not been applied to KRR, except for a theoretical justification. The purpose method of using the circulant matrix in our method is different from previous methods.

Proposed Algorithm Implementation Process

Despite recent advances in KCF tracking algorithms, researchers have paid less attention to the learning section of KCF tracking algorithms. Careful design of this part can solve some challenges, such as drifting and speed issues. Therefore, we propose a new method to improve the learning section of the KCF tracking algorithm. First, we take a look into the background of the learning section based on Kernel ridge regression (KRR).

The classical version of the KRR is well-known for solving complex statistical calculations based on the Hilbert transform. The goal of the KRR calculation is to produce an optimal approximation in the data set $\{(\boldsymbol{x_i},\boldsymbol{y_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}$ using the regression model. The mathematical expectation function $\mathbb{E}$ between X and Y denoted by $f^*(x) = \mathbb{E}[Y|X = x]$.

Notice that the KRR method is based on the convex equation as [12]:

$$f = arg\min_{f\in\mathcal{H}} \left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 + \lambda_n\|f\|_{\mathcal{H}}^2\right\} \quad (12)$$

The finite dimensions of n are used for optimizing the convex equations as:

$$\alpha = arg\min_{\alpha\in\mathbb{R}^n} \left\{\frac{1}{2}\alpha^T K^2\alpha - \alpha^T\frac{K_y}{\sqrt{n}} + \lambda_n\alpha^T K\alpha\right\} \quad (13)$$

$$f(\cdot) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\alpha_i k(.,x_i) \quad (14)$$

where $\alpha, f(\cdot), \lambda_n$ and K are the quadratic program, estimate function, regularization parameter and kernel function, respectively. Superscript of T in (13) is the transpose operator. It should be noted that the temporal and memory complexities are from the order of $\mathcal{O}(n^3)$

and $\mathcal{O}(n^2)$, respectively, leading to an increase in computational complexity and memory and reducing the efficiency of the KRR algorithm.

A. Sketch Kernel Ridge Regression

Adding the circulant matrix to the KRR algorithm creates a new feature. Although the dimensions of the circulant matrix are small, it still needs to be made smaller to reduce the complex calculation and corresponding time. Since the large dimensions of the circulant matrix highly impact the computational speed, it motivated us to propose a sketch circulant matrix in video sequence trackers. This proposed method is called sketched KRR. It should be noted that the sketched KRR (SHKRR) has been previously introduced in statistical calculations [11]. However, this method is not employed in video sequence tracking yet. The SHKRR reduces the size of the kernel matrix and the speed of numerical calculations. In the SHKCF method, the sketched matrix is displayed by $S \in \mathbb{R}^{m\times n}$. This matrix S is extracted from the pattern vector of X. The kernel matrix is sketched, because the dimensions of the kernel matrix are converted from the order of $n \times n$ to the order of $m \times n$ ($m \ll n$). Generally, (17) has been applied to the SHKRR method [11]. Accordingly, the original KRR method is converted to the estimation of sketch kernel ridge regression (SKRR) by the sketch matrix. Therefore, we add (15) to the learning section of the KCF method and implement it in the Matlab platform, particularly for object tracking. This process is referred to as SHKCF in this paper.

$$\alpha' = arg \min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{2} \alpha'^T (SK)(KS^T)\alpha' - \alpha'^T \frac{SK_y}{\sqrt{n}} + \lambda_n \alpha'^T SKS^T \alpha' \right\} \quad (15)$$

It should be noted that this equation is a quadratic sketched program with m dimensions, in which the equations of $(SK^2S^T and\ SKS^T)$ operate as input m dimensional matrices and the equation SK_y is m dimensional vector. To improve the computational, the sketched kernel matrix $SK = [SK_1, \ldots, SK_n]$ in the input is calculated in way that parallelization technique is employed across its columns. In addition, this sketching idea can be extended to other kernel approaches regarding other loss functions that Characterizing of its properties is an interesting research subject for future works [44].

In a sliding window technique, the (16) can be evaluated on total of the sub-windows for fast detecting. However, to compute total of the responses simultaneously, one can exploit the circular technique [1].

$$f'(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (S^T \alpha')_i k(.,x_i) \quad (16)$$

with

$$\alpha' = (SK^2S^T + 2\lambda_n SKS^T)^{-1} SK_y \quad (17)$$

where K and I are the kernel matrix with elements $K_{ij} = k(x_i;\ x_j)$, and the identity matrix, respectively. It should be noted that parameter α', as the estimation function, is needed for solving the KCF filter (H) in (1). The main difference between (6) and (13) is that the sketch matrix is integrated into (13). As a result, we will explain how to calculate the sketch matrix in detail for implementing the KCF method in object tracking (OT), and then we will explain the proposed algorithm in a flowchart.

The key to SKRR success is building the effective Sketch matrix in the tracking. Although many methods have been implemented to improve tracking through KRR upgrades, the Sketch circulant matrix has not yet been used in the tracking.

B. Sketch Kernel Ridge Regression Algorithm

As already mentioned, the main part of the proposed method is the sketched matrix (S matrix) in (11)) which is added to the learning section of the KCF tracking.

$$S = \frac{1}{\sqrt{m}} DCQ \quad s \in \mathbb{R}^{m\times n}, D \in \mathbb{R}^{m\times m}, Q \in \mathbb{R}^{m\times n} \quad (18)$$

where D is Rademacher mapping between +1 or -1 which is similar to the behavior of the random diagonal matrix with a probability of ½. m is the dimension of the sketch matrix. Q represents the sampling matrix which consists of a random subset of m rows from the $n \times n$ identity matrix.

Proposed algorithm: sketch KRR employing Circulant Matrix

Input: Data set $\{(x_i.\, y_i)\}_{i=1}^{n}$

Output: α'

1- Initialization: kernel parameters, sketched matrix m, the regularization parameter $\lambda_n \geq \frac{\sqrt{\log(n)}}{2n}$.
2- Construction of a random diagonal matrix $D \in \mathbb{R}^{m\times m}$, regarding the Rademacher variable
3- Construction of a circulant matrix $C \in \mathbb{R}^{m\times m}$, whose inputs in the first column are obtained from the standard normal distribution.
4- Construction of a variable kernel matrix $K' \in \mathbb{R}^{m\times n}$, which is based on the data sample of the Q matrix in Eq. (18).
5- Calculate $SK \in \mathbb{R}^{m\times n}$ using FFT, $S(SK)^T$.

Calculate $\alpha' = (SK^2S^T + 2\lambda_n SKS^T)^{-1} SK_y$.

The Circulant Kernel Ridge Regression (CKRR) shows our sketch approach. The Gauss kernel is a typical and representative kernel, so in this paper we mainly use Gaussian kernel for experiments. The kernel function (7) is used for all types of feature representations. In the function (9), there are two parameters need to be determined in advance, i.e., σ and λ_n, the kernel

parameter and regularization parameter, respectively. That the λ_n in our approach meets $\lambda_n = \frac{\sqrt{\log(n)}}{2n}$.

Implementation of the SHKCF tracker is simple, i.e. there is no type of heuristic methods for motion modeling or failure detection. Using video sequence dataset SHKCF tracker can train a model at the object's initial position in the first frame of the video sequence. To obtain some context, the searching window of SHKCF tracker is considered larger than the size of object. SHKCF tracker uses the previous position of the object over the window for each new frame. Then, the position of the object is updated to the one which yielded the greatest possible value. To obtain SHKCF tracker with some memory, SHKCF tracker train a new model in its new position. Finally, this tracker interpolates the value of α' linearly with its values which are updated with the following equations of linear interpolation [57].

$$\alpha'^{t} = (1-\theta) * \alpha'^{t-1} + \theta * \alpha' \quad (18)$$

$$X'^{t} = (1-\theta) * X'^{t-1} + \theta * X' \quad (19)$$

where α', X', t and θ denote the kernelized regularized Ridge regression, the object appearance, the t-th frame and the learning rate, respectively. In fact, this updating strategy can work well when changing of object appearance is very slow or there is no occlusion. To solve this problem, two indicators are introduced for evaluating whether the target is in occlusion challenge and tune the learning rate adaptively. If the object is in occlusion challenge the learning rate is reducing; if else the learning rate is fixed. These indicators are Peak-to-Sidelobe Ratio (PSR) [23] and appearance similarity(d). then we tune the learning rate of $\theta = \gamma * \theta_{in}$, if $d \leq 0.22$ and $PSR \leq 30$, $\theta = \theta_{in}$, otherwise. where γ and θ_{in} are the relative ratio to decrease the learning rate and the initialization value, respectively [57].

Fig. 1 shows the implementation steps of the proposed algorithm based on the sketch kernel. As can be seen in the proposed flowchart, a grayscale image sequence is first received to extract the object features. It should be noted that the coded color features of the object in grayscale image sequence are achieved by replicating the gray-scale image sequence into the its red, green and blue types and then extracting the object feature on this image as usual. The video sequence features are extracted by the HOG method. The circulant matrix is then generated and formulated by the first sequence patch of input data. A circulant matrix is used in the ridge regression kernel evaluated by the normal Gaussian distribution.

In the proposed algorithm, the coefficient α' is trained by calculating the sketched kernel correlation filter. Finally, the evaluation is performed to confirm the correct choice of the object. However, if the target is lost, the feature extraction operation is performed again by updating the learning section of the algorithm for the current frame. If the error is less than the specified threshold, the evaluation of the tracking operation is correctly determined, and the next frame is extracted.

As already mentioned, the optimal value of α' is trained with the pattern X. Besides, the coefficient α' should be updated in each sequence of the flowchart (Fig. 1) [19], [35], [36]. The object in the next frame is estimated using the H filter in the search area of the current frame. In other words, the function $f(z;H) = H^T\phi(z)$ is applied to the search area, where z is the evaluation data and the mapping ϕ is the latest updated model of the z.

If the partial mapping ϕ is nonlinear, it can easily estimate several features of the object as a function $f(z;\alpha') = \alpha^T\Phi_{\tilde{x}}\phi(z)$. It should be noted that linear mapping cannot estimate multiple features. According to the element-wise learning algorithm, the mapping z is calculated by considering the circular shifts element by element. Besides, the filter responses in the frequency domain are calculated as a function $f'(z;\alpha') = (k'^{\tilde{x}z} \odot \alpha')$, where $k'^{\tilde{x}z}$ is the estimate of the kernel matrix in the FFT of $\Phi_{\tilde{x}}{\Phi_z}^T$ [33], [35]. Deferent form the earlier type SHKCF tracker [29], it is developed to deal with multiple channels, as input arrays' third dimension. We implement SHKCF tracker by three functions: train (13), detect (14), and kernel correlation (7), which is demonstrated in Fig. 1.

Evaluating the Proposed Method with other Kernel Tracking Algorithms

For a fair comparison, some recent kernel-based video sequence tracking algorithms [29]-[34] are compared to the proposed algorithm. They have been evaluated on the OTB100/50 and VOT2016 challenges video sequence dataset and implemented by the same system (CPU: Core i7 RAM: 8GB, GPU: 4G). Fig. 2 shows the conventional challenges of OTB100/50 object tracking [7], including motion blur (MB), scale variation (SV), out of plane rotation (OPR), illumination variation (IV), occlusion (OCC), in of plane rotation (IPR), out of view (OV), background clutter (BC), low resolution (LR), fast motion (FM), deformation (DEF). For more clarity, three different kinds of one pass evaluation (OPE) are employed: 1- object accuracy, 2- object overlap and 3- qualitative comparison of algorithms. Note that OPE is to evaluate the threshold error of the object location and also overlap with ground truth.

Using a Gaussian kernel, we propose and implement a new object tracker based on the sketch kernelized correlation filter. Then we experimentally test more variants which works on HOG features with 4×4 cell size pixels, in particular variant [42], [53]. It should be noted that we employed adaptation rate (θ_{in}) = 0.02, γ = 0.1,

spatial bandwidth is $S = \frac{1}{\sqrt{m}} DCQ$ (that $m = 1.25\sqrt{\log(n)} \cong 2.17$, $n = 1024$), feature bandwidth $\sigma = 0.5$, and regularization $\lambda = \frac{\sqrt{\log(n)}}{2n} = 8.47 \times 10^{-4}$ [44], in our SHKCF tracker.

A. Evaluation Methods

To examination the tracker's robustness the OPE technique is employed. In this way, The OPE evaluation operates on the object trackers just once in a video sequence. To analyze the trackers' efficiency, plots of precision and success are shown in Fig. 3, Fig. 4 and Fig. 5. The threshold value $t0$ differs in range between 0 and 1 for generating result of curves in the success plots. The threshold value of the success rate is bounded to 0.5 for evaluation process. In the other hand, the Euclidean distance is calculated between the ground truth- and estimated-centers in precision plots, measures center location error (CLE) [12], [31], as follows:

$$\mathrm{CLE} = \upsilon_{gp} = \sqrt{(l_g - l_p)^2 + (r_g - r_p)^2}, \tag{20}$$

where $(lg;\ rg)$ and $(lp;\ rp)$ are the ground truth center location and the predicted center location of the object in a frame, prospectively. During tracking, the average error metric cannot be used to accurately measure the tracking performance, as the tracker can lose the actual object's location and the estimated location can be random. Instead, it may be a better performance metric to use the percentage of frames whose estimated location is within the specified threshold distance from the ground truth.

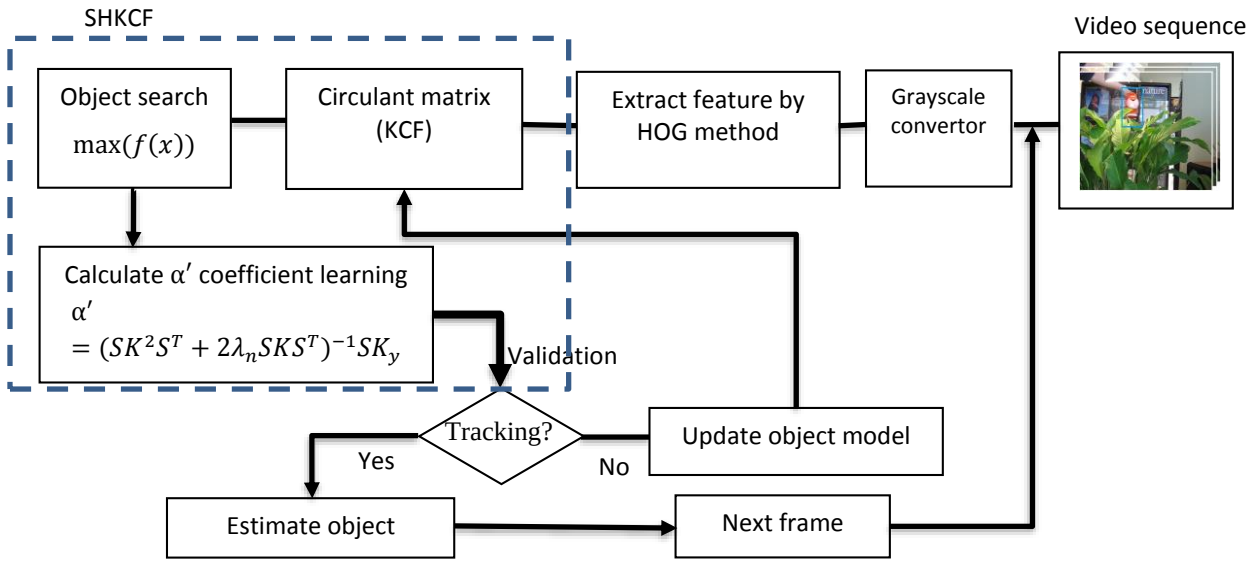


Fig. 1: Flowchart of the proposed tracking algorithm (SHKCF).

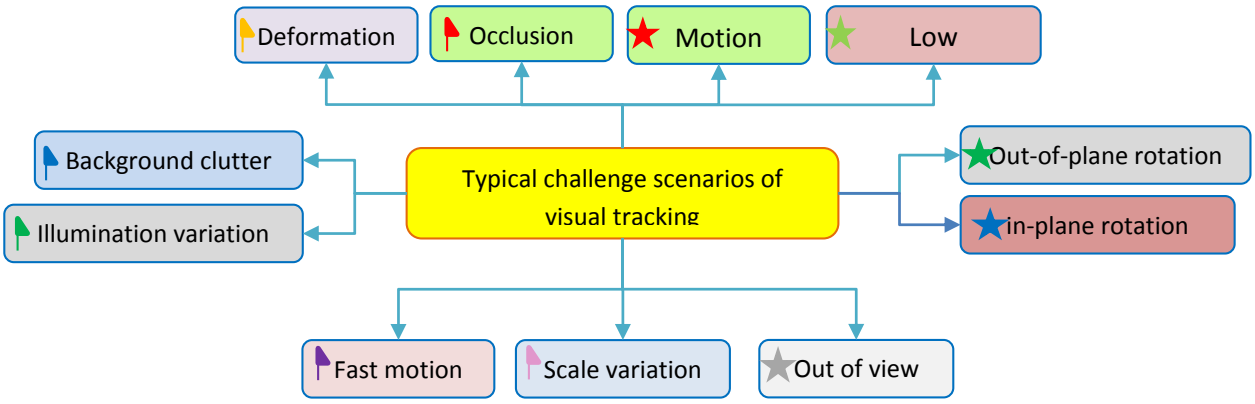


Fig. 2: Conventional challenges of video sequences for goal tracking, from [7].

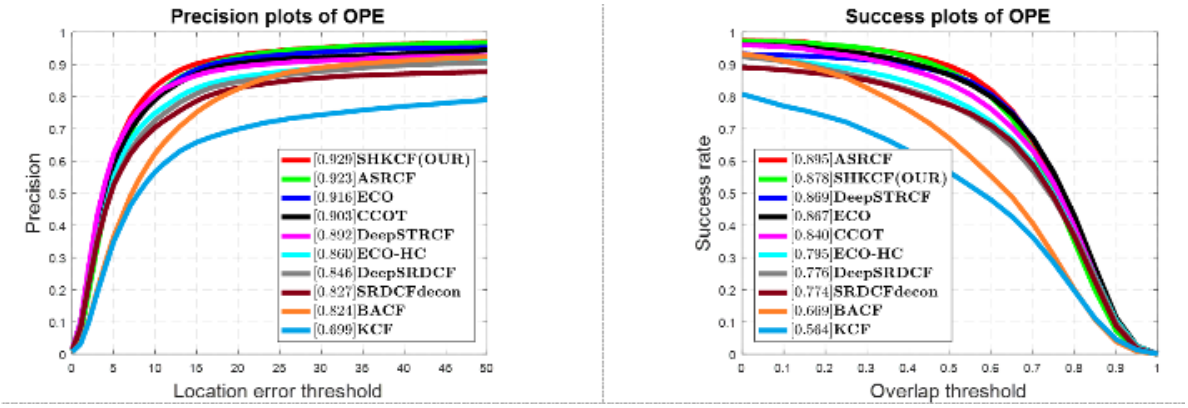


Fig. 3: Success plots (right) and precision plots (left) of our SHKCF tracker against the other ones (ECO [32], CCOT [35], ECO_HC, DeepSRDCF [52], SRDCFdecon [31], KCF [1], ASRCF [59], BACF [50] and DeepSTRCF [52])) on the OTB100 benchmark video sequence.

$$z = \frac{\sum_{n=1}^{N} \chi(v_{gp})}{N} * 100, \tag{21}$$

$$\chi(v_{gp}) = \begin{cases} 1 \ if \ v_{gp}^{n} \leq v_{th} \\ 0 \quad otherwise \end{cases}, \tag{22}$$

where N is all number of frames. Legends in the plots of precision demonstrate that precision regarding to a threshold of $v_{th} = 20$ pixels. Due to error of object center location just measures pixel difference, precision cannot produce a clear picture of estimated object shape and size. Thus, success plots have been employed as a more robustness measurement. In this way, an overlap score (OS) is computed between the ground truth- and the estimated-bounding box, which is based on area under the curve (AUC) [12], [31], as follows:

$$AUC = o_w = \frac{area(|u_t \cap u_g|)}{area(|u_t \cup u_g|)}, \tag{23}$$

where u_t , u_g , |.| ∩ and ∪ are the object bounding box, the ground-truth bounding box, the number of pixels, intersection and union of two regions, respectively. The overlap score is employed to demonstrate whether an object tracking algorithm has been successfully tracked an object in the frame. To demonstrate the successful frames o_w score should have more value than a threshold. Similar to precision, another performance metric is considered for computing of the overlap score percentage, as follows:

$$w = \frac{\sum_{n=1}^{N} \Gamma(o_w^i)}{N} * 100, \tag{24}$$

$$\Gamma(o_w^i) = \begin{cases} 1 \ if \ o_w^i \leq t_0 \\ 0 \quad otherwise \end{cases}, \tag{25}$$

where N and t_0 are all number of frames and the overlap score threshold, respectively.

B. *Quantitative Evaluation*

The overall success performance and the precision of all the trackers over OTB100 are plotted in Fig. 3, where overlap precision (OP) metric is applied by computing the bounding box overlaps greater than 0.5 in a video sequence. In addition, we provide the overlap success plots containing the OP metric over a range of thresholds. We compare SHKCF tracker with 8 state-of-the-art trackers and one base tracker, including trackers (i.e. ECO [32], CCOT [35], ECO_HC, DeepSRDCF [52], SRDCFdecon [31], KCF [1], ASRCF [59], BACF [50] and DeepSTRCF [52])). It should be noted that for fair comparison, we use the publicly available codes or results provided by the authors. As can be seen, in Fig. 3 the result of precision plots of our SHKCF tracker have better than other trackers based on correlation filters and the result of success performance of our SHKCF tracker ranks second after ASRCF tracker with a difference of 0.017. To decrease the drifting problem, the distribution of correlation response is modeled in a sketch optimization framework by the SHKCF algorithm, making the object location in each frame more accurate.

For better understanding, overall comprehensive evaluations of nine top trackers are summarized in Table 1. Table 1 demonstrates that our SHKCF tracker has obtained the best results between three kernel correlation filter based trackers and also other type trackers. Comparing to KCF_DP [1], the SHKCF tracker gets a 32.9%, 55.7% and 78.4% improvement for AUC score, OP score and speed, respectively. Comparing to ASRCF [58], the SHKCF tracker achieves a 0.65%and 6.4% improvement for AUC score and speed, respectively. Also, the overall comprehensive evaluation results show that SHKCF promotes the KCF's performance [11] which

employee the identical scale strategy and features as our object tracker.

Table 1: Overall comprehensive evaluations of our SHKCF and other trackers

Tracker	Mean OP (CLE)	AUC (OS)
SHKCF(OUR+ HOG)	*0.878*	**0.929**
ASRCF	**0.895**	*0.923*
ECO	0.867	0.916
C-COT	0.840	0.903
DeepSTRCF	0.776	0.892
ECO-HC	0.795	0.860
DeepSRDCF	0.776	0.846
SRDCFdecon	0.774	0.827
BACF	0.669	0.824
KCF_DP	0.564	0.699

Table 2 and Table 3 demonstrate OPE partial evaluation of target accuracy based on OTB100 and OTB50 datasets. Basically, the output is based on the area under the curve (AUC).

The evaluation accuracy of the tracking algorithm is between zero and one.

If the output is closer to one, more similarity is obtained between the actual and estimated location of the object. the evaluation was performed in 1500 different conventional video sequences, as OTB100 and OTB50.

Also, the output is an average number of 100 video sequences (OTB100) in Table 2 and Table 3. Ranking in Table 2 and Table 3 is depicted by green/Bold, red/Italic and blue/underline for first, second and third rank, respectively.

In addition, we demonstrate the tracking speed (FPS) comparison on OTB-2015 dataset in

Table 4 One can see that SHKCF (HOGCN) runs at 22.1 FPS. SHKCF (HOG) using HOG feature performs even faster and obtains a real-time speed of 35.4 FPS, which is 1.6× and 1.13× faster than BACF and STRCF tracker, respectively.

As shown in Table 2, the proposed algorithm wins in IPR, BC and FM challenges, Location error overall and Precision OPE (fps). The results of Table 2 demonstrate that the SHKCF method is a good choice considering object tracking accuracy since it gains one of the best ranks in most challenges.

The results of Table 3 demonstrate that the SHKCF method a lower performance in OCC and MB challenges than modern trackers. Fig. 4 and Fig 5 demonstrate the precision and success plots of the proposed method and other SOTA methods, respectively.

The ground-truth location in the first frame acts as an initial value for evaluating the test sequence.

It is worth mentioning that the algorithms are ranked based on AUC.

Table 2: OPE partial evaluation of object accuracy based on AUC calculations

Tracker	illumination Variation [36][36]	Out-of-Plane [63]	Scale Variation [64]	Motion Blur [30]	Occlusion [48]	In-plane-rotation [52]	Out-of-view [15]	Background Clutter[31]	Low resolution [9]	Fast motion [39]	Deformation [44]
SHKCF(OUR)	***0.913***	0.915	***0.905***	0.867	0.889	**0.910**	0.834	**0.943**	0.942	**0.907**	***0.916***
ASRCF	**0.924**	**0.929**	**0.906**	0.881	0.909	***0.897***	**0.918**	0.929	**0.999**	***0.902***	**0.917**
ECO	0.906	***0.918***	*0.892*	**0.888**	**0.927**	0.887	***0.914***	***0.936***	0.882	0.897	0.869
C-COT	0.869	0.903	0.885	***0.887***	***0.921***	0.862	0.890	0.862	***0.977***	0.896	0.871
DeepSTRCF	0.829	0.884	0.871	0.833	0.881	0.830	0.839	0.846	0.881	0.844	0.861
ECO-HC	0.811	0.845	0.832	0.798	0.860	0.802	0.825	0.846	0.888	0.840	0.813
DeepSRDCF	0.766	0.840	0.825	0.817	0.820	0.817	0.787	0.820	0.847	0.831	0.782
SRDCFdecon	0.825	0.803	0.815	0.809	0.779	0.774	0.654	0.839	0.747	0.787	0.754
BACF	0.780	0.811	0.802	0.763	0.758	0.820	0.667	0.771	0.925	0.781	0.805
KCF	0.725	0.689	0.647	0.601	0.642	0.702	0.526	0.715	0.700	0.637	0.629

Table 3: Partial evaluation of OPE object overlap

Tracker	illumination Variation [36]	Out-of-Plane [63]	Scale Variation [64]	Motion Blur [30]	Occlusion [48]	In-plane-rotation [52]	Out-of-view [15]	Background Clutter [31]	Low resolution [9]	Fast motion [39]	Deformation [44]
SHKCF(OUR)	0.867	***0.868***	***0.844***	0.860	0.847	**0.847**	0.794	***0.883***	***0.807***	0.857	***0.857***
ASRCF	**0.900**	**0.889**	**0.858**	***0.875***	**0.882**	**0.847**	**0.856**	**0.907**	**0.808**	**0.874**	**0.868**
ECO	***0.868***	0.863	0.841	**0.881**	***0.865***	***0.829***	***0.800***	0.853	0.717	***0.863***	0.831
C-COT	0.815	0.823	0.817	0.862	0.861	0.775	0.791	0.779	0.738	0.838	0.803
DeepSTRCF	0.811	0.864	0.841	0.833	0.856	0.802	0.789	0.812	0.802	0.832	0.829
ECO-HC	0.767	0.764	0.749	0.775	0.787	0.715	0.740	0.792	0.607	0.785	0.764
DeepSRDCF	0.717	0.753	0.754	0.781	0.736	0.734	0.676	0.724	0.713	0.777	0.708
SRDCFdecon	0.764	0.738	0.765	0.796	0.736	0.715	0.663	0.761	0.679	0.755	0.693
BACF	0.621	0.673	0.560	0.728	0.618	0.678	0.617	0.709	0.341	0.684	0.654
KCF	0.560	0.555	0.449	0.558	0.520	0.570	0.492	0.630	0.304	0.557	0.527

Table 4: The FPS results of trackers on OTB-2015. The best three results are shown in **green**, blue and ***red*** fonts, respectively.

	ECO_HC	SRDCF	SRDCFDecon	KCF_DP	STRCF	DeepSRDCF	CCOT	BACF	SHKCF(HOGCN)	SHKCF (HOG)
FPS	15.6	5.8	2.0	16.7	***31.5***	5.3	0.3	26.7	22.1	**35.4**

The number in the parentheses specifies how many sequences are used in each challenging scenario. According to the precision of OPE in Fig. 4, our SHKCF algorithm finds the best result results in some challenging scenarios (MB, OCC and OV). Moreover, the experimental result of Fig. 4 shows that our SHKCF method achieves second and third rank in other challenging scenarios. The horizontal and vertical axis in each precision plot of OPE indicates the location error and overlap threshold of the OT algorithm, respectively (Fig. 4 and Fig. 5). Comparing results between the proposed method and the top ten related works demonstrate that our method has a close or even better performance than other methods. Note that all methods are implemented using the same platform challenges and processor. The final result of the AUC calculation is presented at the bottom right of each precision plot in Fig. 4.

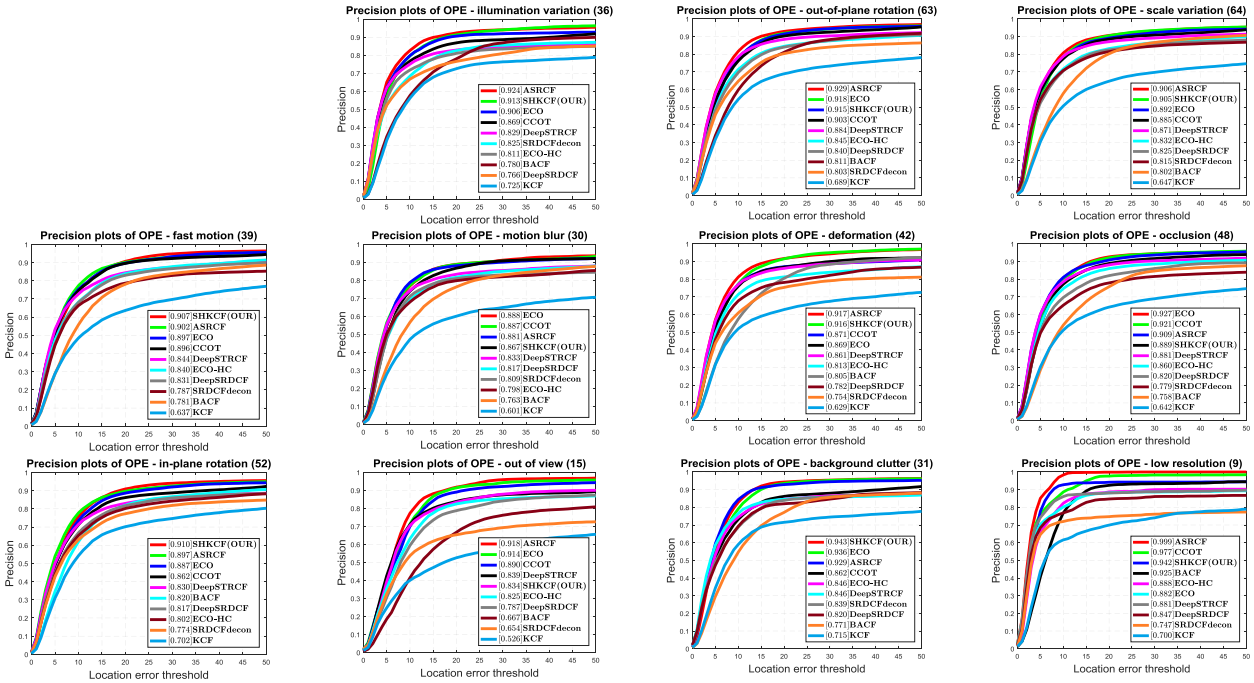


Fig. 4: Precision distance plot of our SHKCF tracker and other trackers (ECO [32], CCOT [35], ECO_HC, DeepSRDCF [52], SRDCFdecon [31], KCF [1], ASRCF [59], BACF [50] and DeepSTRCF [52])) on OTB100 [12] benchmark on eleven different challenges (IV, OPR, SV, FM, MB, Def, OCC, IPR, OV, BC and LR). The legend has score at a threshold of 20 pixels for each object tracker.

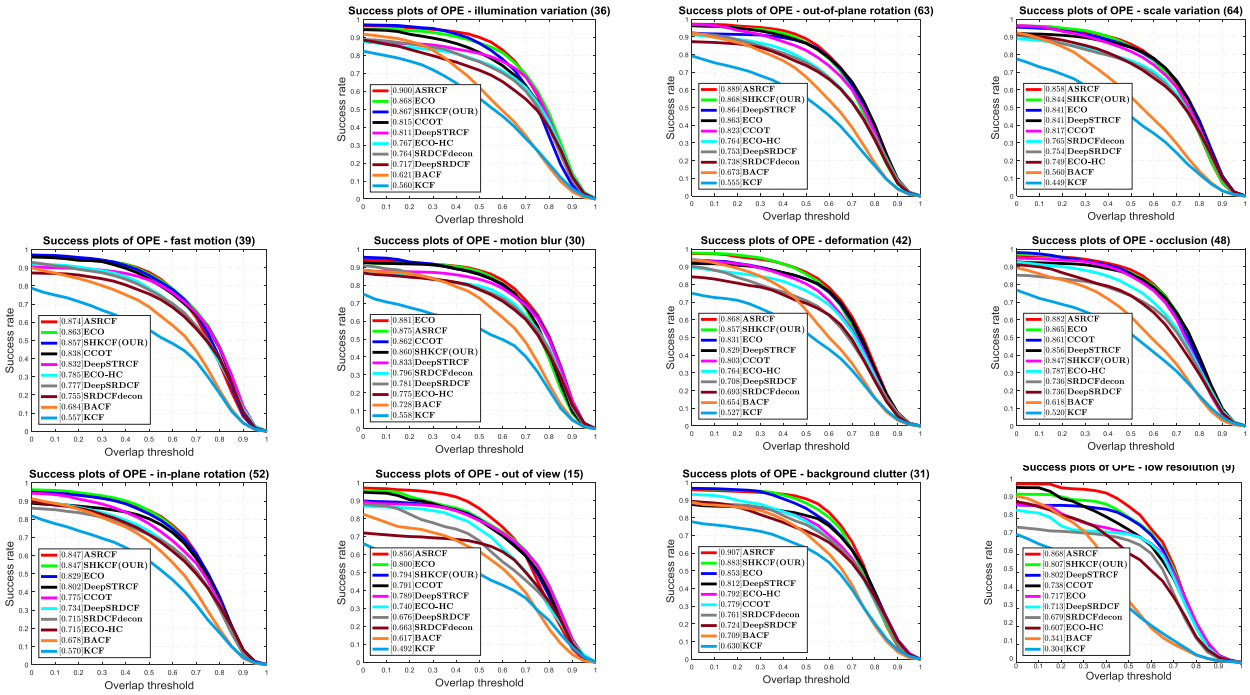


Fig. 5: Overlap success plots of our SHKCF tracker and other trackers (ECO [32], CCOT [35], ECO_HC, DeepSRDCF [52], SRDCFdecon [31], KCF [1], ASRCF [59], BACF [50] and DeepSTRCF [52])) on OTB100 [12] benchmark on eleven different challenges (IV, OPR, SV, FM, MB, Def, OCC, IPR, OV, BC and LR). The legend has score at a threshold 0.5 pixels for each object tracker.

C. Qualitative Evaluation

For more clarity, Fig. 6 shows the result of the qualitative comparisons evaluated between SOTA works (e.g., ECO, CCOT, ECO_HC, DeepSRDCF, SRDCFdecon, KCF, ASRCF, BACF, DeepSTRCF) and SHKCF. In order to cover overall the tracking challenges, random sequences are selected as shown in Fig. 6, which is including Soccer, Matrix, Ironman, Basketball, Bolt, Bird, Deer, Biker and Panda. These samples include several available challenges of the OTB 100 benchmark.

Soccer video sequence sample:

In this sample with 383 frames evaluates on SV, OCC, MB, IPR, OPR and BC challenges. According to experimentally result Fig. 6, in frame 30, all trackers were able to track the target correctly, it should be noted that in its 70th frame SV, MB and IPR challenges are evaluated, and the tracking drift is observed in the evaluation of KCF and DeepSRDCF trackers. In frame 110, four trackers (KCF, BACF, ARSCF and DeepSTRCF) completely lost the target, and five trackers (ECO, C_COT, ECO_HC, SRDCFDecon and DeepSRDCF) partially overlap the target, and only the proposed tracker of SHKCF tracker succeeds in tracking the target. Farther more in frame 345, KCF and DeepSRTCF trackers completely lost the target and the other six trackers and our tracker found the target position and overlapped the target.

Matrix video sequence sample:

In this sample with 91 frames evaluates on IV, SV, OCC, FM, IPR, OPR and BC challenges.

According to Fig. 6, in frame 15, DeepSRDCF and C_COT trackers have partial overlap and the other trackers and our tracker have full overlap. At frame 40, except for the ECO and C_COT trackers which have partial overlap, the all other trackers lost the target. In the next frame of matrix sample, except the ECO and C_COT trackers which have complete overlap, the all other trackers lost the target. In this example, the proposed tracker lost the target from frame 40 onwards, and the tracker could not correct itself in the next frames.

In this sample with 91 frames evaluates on IV, SV, OCC, FM, IPR, OPR and BC challenges. According to Fig. 6, in frame 15, DeepSRDCF and C_COT trackers have partial overlap and the other trackers and our tracker have full overlap. At frame 40, except for the ECO and C_COT trackers which have partial overlap, the all other trackers lost the target. In the next frame of matrix sample, except the ECO and C_COT trackers which have complete overlap, the all other trackers lost the target. In this example, the proposed tracker lost the target from frame 40 onwards, and the tracker could not correct itself in the next frames.

Ironman video sequence sample:

In this sample with 157 frames evaluates on IV, OCC, FM, IPR, OPR, OV and BC challenges. According to Fig. 6, in frame 36, the BACF and DeepSRDCF trackers the target is lost the target, the other trackers have partial overlap, and the proposed tracker has the most overlap with the target. At frame 54, four trackers (BACF, KCF, DeepSRDCF and SRDCFdeecon) have lost the target, and one tracker

(DeepSRTCF) has partial overlap and the other trackers, including the proposed SHKCF tracker, have full target overlap.

Basketball video sequence sample:

In this sample with 716 frames evaluates on IV, OCC, FM, IPR, OPR, OV and BC challenges. According to Fig. 6, in frame 30, only the SRDCFdecon tracker lost the target. At frame 650, the ECO tracker with partial occlusion, the target window is larger than the ground-truth and does not overlap completely. At frame 717, the ECO tracker misses the target completely, the SRDCFdecon tracker has partial overlap, and the other trackers track the target correctly.

Bolt video sequence sample:

In these sample with 341 frames evaluates on OCC, OPR and BC challenges. According to Fig. 6, except for the SRDCFdecon tracker, the other trackers found the target correctly.

Bird video sequence sample:

In this sample with 716 frames evaluates on OCC, DEF, FM, IPR, OPR and OV challenges. According to Fig. 6, in frame 20, except for the KCF tracker which has a partial target overlap, the other trackers found the target correctly. At frame 123, three SRDCFdecon, KCF, ECO_HC trackers lost the target. At frame 182, the proposed tracker found the target completely, and the two trackers have partial overlap, and the other trackers lost the target. In frame 391, only two SHKCF and ASRCF trackers found the target correctly.

Deer video sequence sample:

In this sample with 62 frames evaluates on MB, FM, IPR and BC challenges. According to Fig. 6, in frame 20, all trackers track the target correctly. In frame 30, three trackers (DeepSRDCF, ECO and C_COT) lost the target.

At frame 40, only one tracker (C_COT) lost the target and the other trackers found the target correctly. At frame 40, all trackers found the target correctly.

Biker video sequence sample:

In this sample with 133 frames evaluates on OCC, MB, FM, OPR, OV and LR challenges. According to Fig. 6, in frame 68, three trackers (DeepSRDCF, BACF and C_COT) lost the target and other trackers, including the proposed SHKCF tracker, follow the target correctly. In frames 80, 84, and 140, the proposed tracker and five other trackers (SHKCF, ECO_HC, DeepSRTCF, KCF, ECO and SRDCFDecon) found the tracker with proper overlap.

Panda video sequence sample:

In this sample with 991 frames evaluates on DEF, IPR, OPR, OV and LR challenges. According to Fig. 6, in frame 110, the trackers are tracking the target correctly. At frame 213, only one tracker (SRDCFDecon) has missed the target, and other trackers, including the proposed tracker, are tracking the target. At frame 345, the proposed tracker and seven other trackers found the target with proper overlap. At frame 645, three trackers (C_COT, DeepSRDCFand SRDCDecon) lost the target and the proposed tracker found the target correctly with other modern trackers.

The SHKCF, ASRCF, ECO, the DeepSRDCF, the CCOT and the SRDCFdecon succeeded in object tracking in a clean environment. To highlight the strange of our tracker against other trackers, quality evaluation in the illumination variation of the object, clearly shows that our tracker can only track the object in the "Matrix" challenges. These quantity output results corroborate that our method has better experimental results than other works. As mentioned before, we employ the same parameters and the protocol generated in the OTB100 for all video sequences.

Table 5: Evaluation on VOT2016 benchmark by expected average ovelap (EAO), Accuracy and robustness

	ECO	SRDCF	BACF	SRDCFDecon	ECO_HC	DeepSTRCF	SHKCF (OUR)
EAO	**0.369**	0.249	0.221	0.259	***0.329***	0.318	0.321
Accuracy	0.52	0.51	**0.58**	0.51	0.54	***0.56***	0.54
Robustness	**0.78**	1.61	1.80	1.49	1.13	0.96	***0.89***

A. The VOT-2016 Benchmark

We demonstrate the results on VOT2016 [21] benchmark, the VOT2016 benchmark consists of 60 video sequences. We evaluate the trackers of expected average overlap (EAO), accuracy and robustness [31].

The EAO measures the average without-reset overlap of a tracker over several video sequences. The accuracy computes the average overlap ratio between the predicted and the ground-truth box. And the robustness averages the number of tracking failures on the video sequence.

Fig. 6: Qualitative analysis of trackers SHKCF(OUR), ECO [32], CCOT [35], ECO_HC, DeepSRDCF [52], SRDCFdecon [31], KCF [1], ASRCF [59], BACF [50] and DeepSTRCF [52] over OTB100 [12] benchmark on eleven challenging sequences (from up to down Soccer, Matrix, Ironman, Basketball, Bolt, Bird, Deer, Biker and Panda respectively).

Results and Discussion

We compare SHKCF with state-of-the-art trackers, including ECO [32], SRDCF [52], BACF [50], SRDCFDecon [31], ECO_HC [32] and DeepSTRCF [52]. Table 5 demonstrates the results of different trackers on VOT-2016 dataset.

We can see from Table 5 that SHKCF performs significantly better than the DeepSTRCF, BACF and SRDCF methods in terms of the EAO metric. In addition, SHKCF also performs favorably against DeepSTRCF, BACF and SRDCF by a gain of 2.6%, 4.4% and 2.9% in EAO metric, respectively.

Compared to modern trackers, the SHKCF tracker was ranked second and three in robustness and accuracy, respectively.

Conclusion

In this paper, we have proposed, analyzed, and implemented the sketch kernel correlation filter (SHKCF) for object tracking.

The proposed method improves the basic correlation filter trackers. Although the sketch matrix theory was first proposed in regression, this method was not employed in object tracking scenarios which motivated us to employ it in the KCF basic algorithm. We exploit the learning section of the filter by integrating a new parameter α with original KCF trackers.

To speed up learning and detection, the element-wise matrix is trained by a sketch algorithm. The element-wise matrix is developed by a circulant matrix to sketch method.

Experimental results on OTB100 and OTB50 standard challenges such as MB, SV, OP, IV, OCC, IPR, OV, BC, LR, FM demonstrate that the SHKCF algorithm can further develop the original KCF tracker performance compared to most of the state-of-art works. In addition, the SHKCF method provides optimal optimization for scaling and occlusion problems.

Besides, the proposed algorithm shows that it has better accuracy and robustness compared to other trackers based on in-depth training, ECO, CCOT, ECO_HC, DeepSRDCF, SRDCFdecon, KCF, ASRCF, BACF and DeepSTRCF.

Finally, the experimental results of the quality comparison of our SHKCF method with other SOTA works related to the online CFT demonstrate the better performance of the proposed algorithm. In order to developed our SHKCF method for the object tracking, the BACF algorithm may be incorporated to the Sketch coefficient for the learning model, leading to increase the learning speed.

Therefore, in addition to the advantages the BACF algorithm such as efficient background target modeling, the mention boundary effect of these new incorporating method of BACF. and SHKCF can solves the local minimum problem BACF.

This new algorithm may perform better in various challenges.

Author Contributions

M. Yousefzadeh, A. Golmakani, and Gh. Sarbishaei designed the experiments. M. Yousefzadeh collected the data. M. Yousefzadeh carried out the data analysis. M. Yousefzadeh, A. Golmakani, and Gh. Sarbishaei interpreted the results and wrote the manuscript.

Acknowledgment

The author gratefully acknowledges the sadjad university A. Golmakani, gh. Sarbishaei for their work on the original version of this document.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

MB	Motion Blur
SV	Scale Variation
OPR	Out of Plane Rotation
IV	Illumination Variation
OCC	Occlusion
IPR	In of Plane Rotation
OV	Out of View
BC	Background Clutter
LR	Low Resolution
FM	Fast Motion
DEF	Deformation
OPE	One Pass Evaluation
CLE	Center Location Error
AUC	Area Under the Curve
KCF	Kernel Correlation Filter

References

[1] J. F. Henriques, R. Caseiro, P. Martins, J. Batista, "Highspeed tracking with kernelized correlation filters," IEEE Trans. Pattern Anal. Mach. Intell., 37(3): 583–596, 2014.

[2] P. Zhang, S. Yu, J. Xu, Xinge You, X. Jiang, X. Y. Jing, D. Tao, "Robust visual tracking using multi-frame multi-feature joint modeling," IEEE Trans. Circuits Syst. Video Technol., 29(12): 3673–3686, 2018.

[3] B. Chen, P. Li, C. Sun, D. Wang, G. Yang, H. Lu, "Multi attention module for visual tracking," Pattern Recognit., 87: 80–93, 2019.

[4] P. Li, D. Wang, L. Wang, H. Lu, "Deep visual tracking: Review and experimental comparison," Pattern Recognit., 76: 323–338, 2018.

[5] T. Xu, Z. H. Feng, X. J. Wu, J. Kittler, "Learning low-rank and sparse discriminative correlation filters for coarse-to-fine visual object tracking," IEEE Trans, Circuits Syst. Video Technol., 30(10): 3727–3739, 2019.

[6] B. Zhang, Z. Li, X. Cao, Q. Ye, C. Chen, L. Shen, A. Perina, R. Jill, "Output constraint transfer for kernelized correlation filter in tracking," IEEE Trans. Syst. Man Cybern. Syst., 47(4): 693–703, 2016.

[7] H. Wang, H. Xu, Y. Yuan, "High-dimensional expensive multi-objective optimization via additive structure," Intell. Syst. Appl., 14: 200062, 2022.

[8] C. Du, M. Lan, M. Gao, Zh. Dong, H. Yu, Zh. He, "Real-time object tracking via adaptive correlation filters," Sensors, 20(15): 4124, 2020.

[9] M. Y. Abbass, K. C. Kwon, N. Kim, S. A Abdelwahab, F. E. A. El-Samie, A. A. Khalaf, "A survey on online learning for visual tracking," Visual Comput., 37: 993–1014, 2021.

[10] X. Zhang, G. S. Xia, Q. Lu, W. Shen, L. Zhang, "Visual object tracking by correlation filters and online learning," ISPRS J. Photogramm. Remote Sens., 140: 77–89, 2018.

[11] R. Yin, Y. Liu, W. Wang, D. Meng, "Sketch kernel ridge regression using circulant matrix: Algorithm and theory," IEEE Trans. Neural Networks Learn. Syst., 31(9): 3512–3524, 2019.

[12] B. Yan, H. Zhao, D. Wang, H. Lu, X. Yang, "Skimming-perusal tracking: A framework for real-time and robust longterm tracking," in Proc. the IEEE/CVF International Conference on Computer Vision: 2385–2393, 2019.

[13] S. Avidan, "Ensemble tracking," IEEE transactions on pattern analysis and machine intelligence, 29(2):261–271, 2007.

[14] J. Zheng, Bo. Li, P. Tian, G. Luo, "Robust object tracking using valid fragments selection," in Proc. Multi Media Modeling: 22nd International Conference, Part I 22: 738–751, 2016.

[15] H. Yang, S. Qu, "Online hierarchical sparse representation of multifeature for robust object tracking," Comput. Intell. Neurosci., 330-341, 2016.

[16] D. T Nguyen, C. D Nguyen, R. Hargraves, L. A Kurgan, K. J. Cios, "mi-ds: Multiple-instance learning algorithm," IEEE Trans. Cybern., 43(1): 143–154, 2012.

[17] D. Comaniciu, V. Ramesh, P. Meer, "Kernel-based object tracking," IEEE Trans. Pattern Anal. Mach. Intell., 25(5): 564–577, 2003.

[18] M. Danelljan, G. Hager, F. Shahbaz Khan, M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in Proc. the IEEE International Conference on Computer Vision: 4310–4318, 2015.

[19] H. K. Galoogahi, A. Fagg, S. Lucey, "Learning background-aware correlation filters for visual tracking," in Proc. the IEEE International Conference on Computer Vision: 1135– 1143, 2017.

[20] Q. Guo, R. Han, W. Feng, Zh. Chen, L. Wan, "Selective spatial regularization by reinforcement learned decision making for object tracking," IEEE Trans. Image Proc., 29: 2999–3013, 2019.

[21] S. Hadfield, R. Bowden, K. Lebeda, "The visual object tracking vot2016 challenge results," Lect. Notes Comput. Sci., 9914: 777–823, 2016.

[22] M. Danelljan, G. Hager, F. Shahbaz Khan, M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in Proc. the IEEE conference on Computer Vision and Pattern Recognition: 1430–1438, 2016.

[23] M. Danelljan, G. Bhat, F. Shahbaz Khan, M. Felsberg, "Eco: Efficient convolution operators for tracking," in Proc. the IEEE conference On Computer Vision and Pattern Recognition: 6638–6646, 2017.

[24] A. Yilmaz, O. Javed, M. Shah, "Object tracking: A survey," Acm Comput. Surv. (CSUR), 38(4): 13–es, 2006.

[25] X. Mei, H. Ling, Y. Wu, E. Blasch, L. Bai, "Minimum error bounded efficient l1 tracker with occlusion detection," in Proc. CVPR: 1257–1264, 2011.

[26] Sh. Shekhar, N. Hoque, D. K. Bhattacharyya, "Pknnmifs: A parallel knn classifier over an optimal subset of features," Intell. Syst. Appl., 14: 200073, 2022.

[27] D. A. Ross, J. Lim, et al., "Incremental learning for robust visual tracking," Int. J. Comput. Vision, 77(1-3): 125-141, 2008.

[28] Y. Li, J. Zhu, et al., "A scale adaptive kernel correlation filter tracker with feature integration," in Proc. ECCV workshops (2), 8926: 254–265, 2014.

[29] M. Danelljan, G. Häger, F. Khan, M. Felsberg, "Accurate scale estimation for robust visual tracking," in Proc. British Machine Vision Conference, Bmva Press, 2014.

[30] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, R. Pflugfelder, "The visual object tracking vot2015 challenge results," in Proc. the IEEE International Conference on Computer Vision Workshops: 1–23, 2015.

[31] H. Lian, F. Zhang, W. Lu, "Randomized sketches for kernel cca," Neural Networks, 127: 29–37, 2020.

[32] A. Mahalanobis, B. V. Kumar, D. Casasent, "Minimum average correlation energy filters," Appl. Opt., 26(17): 3633– 3640, 1987.

[33] B. V. Kumar, "Minimum-variance synthetic discriminant functions," JOSA A, 3(10): 1579–1584, 1986.

[34] D. S. Bolme, J. R. Beveridge, B. A. Draper, Y. M. Lui, "Visual object tracking using adaptive correlation filters," in Proc. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition: 2544–2550, 2010.

[35] S. Liu, D. Liu, G. Srivastava, D. Połap, M. Woźniak, "Overview of correlation filter based algorithms in object tracking," Complex Intell. Syst., 7: 1895–1917, 2020.

[36] M. Danelljan, A. Robinson, F. Sh. Khan, M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in Proc. Computer Vision–ECCV 2016: 14th European Conference: 472–488, 2016.

[37] T. Liu, G. Wang, Q. Yang, "Real-time part-based visual tracking via adaptive correlation filters," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 4902– 4912, 2015.

[38] C. Ma, X. Yang, Ch. Zhang, M. H. Yang, "Long-term correlation tracking," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 5388–5396, 2015.

[39] D. S. Bolme, J. R. Beveridge, B. A. Draper, Y. M. Lui, "Visual object tracking using adaptive correlation filters," in Proc. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition: 2544–2550, 2010.

[40] G. Turin, "An introduction to matched filters," IRE Trans. Inf. Theory, 6(3): 311–329, 1960.

[41] E. Gundogdu, A. A. Alatan, "Good features to correlate for visual tracking," IEEE Trans. Image Process., 27(5): 2526– 2540, 2018.

[42] J. F. Henriques, R. Caseiro, P. Martins, J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in Proc. Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Part IV 12: 702–715, 2012.

[43] Y. Yang, M. Pilanci, M. J. Wainwright, "Randomized sketches for kernels: Fast and optimal nonparametric regression," arXiv preprint arXiv:1501.06195, 2017.

[44] L. Čehovin, A. Leonardis, M. Kristan, "Visual object tracking performance measures revisited," IEEE Trans. Image Process., 25(3): 1261–1274, 2016.

[45] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., 32(9): 1627–1645, 2009.

[46] Q. He, X. Shao, W. Chen, X. Li, X. Yang, T. Sun, "Adaptive multi-scale tracking target algorithm through drone," IEICE Trans. Commun., 102(10): 1998–2005, 2019.

[47] K. Zhang, J. T Kwok, "Clustered nyström method for large scale manifold learning and dimension reduction," IEEE Trans. Neural Networks, 21(10): 1576–1587, 2010.

[48] Sh. Xuan, Sh. Li, M. Han, X. Wan, G. S. Xia, "Object tracking in satellite videos by improved correlation filters with motion estimations," IEEE Trans. Geosci. Remote Sens., 58(2): 1074–1086, 2019.

[49] M. Ding, W. H. Chen, L. Wei, Y. F. Cao, Z. Y. Zhang, "Visual tracking with online assessment and improved sampling strategy," IEEE Access, 8: 36948–36962, 2020.

[50] J. P. Sun, E. J. Ding, B. Sun, Zh. Y. Liu, K. L. Zhang, "Adaptive kernel correlation filter tracking algorithm in complex scenes," IEEE Access, 8: 208179–208194, 2020.

[51] G. Zhu, J. Wang, Y. Wu, X. Zhang, H. Lu, "Mchog correlation tracking with saliency proposal," in Proc. the AAAI Conference on Artificial Intelligence, 30, 2016.

[52] M. Danelljan, A. Robinson, F. S. Khan, M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in Proc. Computer Vision–ECCV 2016: 14th European Conference, Part V 14: 472–488, 2016.

[53] X. Li, Q. Liu, Zh. He, H. Wang, Ch. Zhang, W. S. Chen, "A multi-view model for visual tracking via correlation filters," Knowledge-Based Systems, 113: 88–99, 2016.

[54] Y. Zhang, L. Wang, J. Qi, D. Wang, M.Feng, H. Lu, "Structured siamese network for real-time visual tracking," in Proc. the European Conference on Computer Vision (ECCV): 351–366, 2018.

[55] X. Hao, J. Huang, F. Qin, X. Zheng, "Multi-label learning with missing features and labels and its application to text categorization," Intell. Syst. Appl., 14: 200086, 2022.

[56] M. Tang, B. Yu, F. Zhang, J. Wang, "High-speed tracking with multi-kernel correlation filters," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 4874–4883, 2018.

[57] P. Dollár, R. Appel, S. Belongie, P. Perona, "Fast feature pyramids for object detection," IEEE Trans. Pattern Anal. Mach. Intell., 36(8): 1532–1545, 2014.

[58] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in Proc. the IEEE/CVF conference on Computer Vision and Pattern Recognition: 4282–4291, 2019.

[59] K. Dai, D. Wang, H. Lu, C. Sun, J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 4670–4679, 2019.

Biographies

Morteza Yousefzadeh He was born in the city of Mahmudabad-Mazandaran in Iran. He studied B.A. in Noshirvani University of Babol and M.A. in Malek Ashtar University. His work experience includes digital electronic systems and image processing. His specific areas of interest include artificial intelligence, image processing, and image tracking algorithms. He is a Ph.D. student in Electrical and Electronic Engineering from Sajjad University of Mashhad.

- Email: morteza.usefzadeh@gmail.com
- ORCID: 0009-0006-2611-1616
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Abbas Golmakani He was born in Mashhad in Iran. He studied B.A. and M.A. in Sharif University of Technology. His work experience includes designing analog, digital and high frequency integrated circuits - IoT systems. His special areas of interest include integrated circuit design, image processing. He studied Ph.D. Electrical and Electronics Engineering at Ferdowsi University of Mashhad.

- Email: golmakani@sadjad.ac.ir
- ORCID: 0009-0006-6061-4380
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://profile.sadjad.ac.ir/golmakani/

Ghazaleh Sarbishaei She was born in Mashhad in Iran. She studied B.A. and M.A. in Ferdowsi University of Mashhad. His work experience includes telecommunication systems and signal processing. His special fields of interest include telecommunication systems and signal processing. She studied Ph.D. Electrical and Telecommunication Engineering at Ferdowsi University of Mashhad.

- Email: gh_sarbisheie@sadjad.ac.ir
- ORCID: 0000-0002-7590-2928
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://profile.sadjad.ac.ir/sarbisheie

How to cite this paper:
M. Yousefzadeh, A. Golmakani, G. Sarbishaei, "Design, analysis, and implementation of a new online object tracking method based on Sketch Kernel Correlation Filter (SHKCF)," J. Electr. Comput. Eng. Innovations, 12(1): 115-132, 2024.

DOI: 10.22061/jecei.2023.10126.680

URL: https://jecei.sru.ac.ir/article_1989.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

The New Family of Adaptive Filter Algorithms for Block-Sparse System Identification

E. Heydari [1], M. Shams Esfand Abadi [1,], S. M. Khademiyan [2]*

[1]*Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.*

[2]*Department of Applied Mathematics, Shahid Rajaee Teacher Training University, Tehran, Iran.*

Article Info

Article History:
Received 22 July 2023
Reviewed 02 August 2023
Revised 16 September 2023
Accepted 11 October 2023

Keywords:
Block sparse
$L_{2,0}$-norm
IMSAF
Selective regressors
Dynamic selection

*Corresponding Author's Email Address: *mshams@sru.ac.ir*

Abstract

Background and Objectives: In order to improve the performance of normalized subband adaptive filter algorithm (NSAF) for identifying the block-sparse (BS) systems, this paper introduces the novel adaptive algorithm which is called BSNSAF. In the following, an improved multiband structured subband adaptive filter (IMSAF) algorithms for BS system identification is also proposed. The BS-IMSAF has faster convergence speed than BS-NSAF. Since the computational complexity of BS-IMSAF is high, the selective regressor (SR) and dynamic selection (DS) approaches are utilized and BS-SR-IMSAF and BS-DS-IMSAF are introduced. Furthermore, the theoretical steady-state performance analysis of the presented algorithms is studied.

Methods: All algorithms are established based on the $L2,0$-norm constraint to the proposed cost function and the method of Lagrange multipliers is used to optimize the cost function.

Results: The good performance of the proposed algorithms is demonstrated through several simulation results in the system identification setup. The algorithms are justified and compared in various scenarios and optimum values of the parameters are obtained. Also, the computational complexity of different algorithms is studied. In addition, the theoretical steady state values of mean square error (MSE) values are compared with simulation values.

Conclusion: The BS-NSAF algorithm has better performance than NSAF for BS system identification. The BSIMSAF algorithm has better convergence speed than BS-NSAF. To reduce the computational complexity, the BS-SR-IMSAF and BS-DSR-IMSAF algorithms are developed. These algorithms have close performance to BS-IMSAF.

Introduction

The least mean squares (LMS) and the normalized LMS (NLMS) algorithms are widely used in many adaptive filter applications [1]-[4]. These algorithms are simple, stable and easy to implement [5]. However, the convergence speed of LMS and NLMS algorithms is significantly deteriorated in case of colored input signals. To improve the convergence speed of these algorithms, different algorithms such as affine projection (AP) and normalized subband adaptive filter (NSAF) algorithms were introduced [6], [7]. Also, the methods of APA and NSAF were combined and improved multiband structured SAF (IMSAF) was proposed in [8], [9]. Since the computational complexity of APA and IMSAF is high, various approaches such as selective regressor (SR) and dynamic selection regressor (DSR) were applied in APA [10], [11] and IMSAF [12]-[14] as well as wavelet transform domain LMS (WTDLMS) [15].

In some applications, the unknown system to be

identified is sparse or block-sparse (BS). It means that the unknown system consists of many zero or near-zero coefficients and a small number of large ones. The typical sparse systems are digital TV transmission channels and echo paths [16]. Also, in satellite-linked or in-door MIMO communications, the impulse response is block sparse. The classical adaptive filter algorithms such as NLMS, APA, NSAF, and IMSAF suffer from poor performance when the impulse response of the unknown system is sparse or block sparse [17].

To solve this problem, the L_0-norm constraint are utilized in the cost function of various adaptive filter algorithms. In [18], the L_0-LMS was presented which shows better performance than LMS and NLMS in sparse system identifications. Also, two types of L_0-APA and L_0-NSAF algorithms were proposed in [6], [7]. The other researches on sparse systems can be found in [20], [23], [24], and [25]. Furthermore, in our recent research, we introduced the L_0-IMSAF algorithm [26]. As we mentioned, there is a special sparse system which is called block-sparse (BS). The impulse response of block-sparse system consists of one or more clusters, wherein a cluster is a gathering of nonzero coefficients. In this situation, the sparse adaptive algorithm such as L_0-LMS doesn't work well. Therefore, the BS-LMS was introduced [27]. The BS-LMS has much better convergence speed than L_0-LMS in BS system identification. In BS-LMS, a penalty of BS, which is mixed $L_{2,0}$-norm of adaptive filter coefficients with equal group partition sizes, is inserted to the cost function of LMS. This approach was successfully extended to proportionate NLMS in [28].

In the present study, the BS-NSAF algorithm is firstly introduced. The BS-NSAF has faster convergence speed than L_0-NSAF in block-sparse systems. Then, to improve the performance of BS-NSAF, the BS-IMSAF is presented. Both algorithms are established based on the $L_{2,0}$-norm constraint to the proposed cost function. In the following, we introduce two new algorithms to reduce the computational complexity of BS-IMSAF. The SR and DSR approaches are extended to BS-IMSAF and BS-SR-IMSAF and BS-DSR-IMSAF are established. In BS-SR-IMSAF, a subset of the input regressors at each subband are optimally selected during the adaptation. The subsets with dynamic number of members from the input regressors (DSR) at each subband are chosen for every iteration in BS-DSR-IMSAF. Furthermore, the theoretical steady-state performance analysis of the proposed algorithms is also studied. Table 1 reviews the classical, sparse, and BS adaptive filter algorithms. The proposed algorithms have been indicated in Table 1. Also, Table 2 compares the cost functions of L_0-LMS, BS-LMS, L_0-NSAF, IMSAF, and proposed BS-NSAF and BS-IMSAF algorithms. In the following, the notations in this table will be illustrated.

Table 1: The L_0-norm constraint adaptive filter algorithms

Algorithm	Algorithm based on Block Sparse
LMS	L_0-LMS [18], BS-LMS [27]
NSAF	ZN-NSAF [19], BS-NSAF *
IMSAF	BS-IMSAF *
SR-IMSAF [12]	BS-SR -IMSAF *
DSR-IMSAF [12]	BS-DSR-IMSAF *

* Proposed in this paper.

Table 2: Review of cost functions

Adaptive filter algorithm	Cost function
L_0-LMS [18]	$J(k) = \lvert e(k)\rvert^2 + \delta \parallel \mathbf{h}(k) \parallel_0$
BS-LMS [27]	$J(k) = \lvert e(k)\rvert^2 + \delta \parallel \mathbf{h}(k) \parallel_{2,0}$
ZN-NSAF-I [19]	$J(n) = \parallel \mathbf{h}(n+1) - \mathbf{h}(n) \parallel^2 + \sum_{i=1}^{N} \lambda_i [d_{i,D}(n) - \mathbf{u}_i^T(n)\mathbf{h}(n+1)]$
ZN-NSAF-II [19]	$J(n) = \frac{1}{2}\sum_{i=1}^{N} \left(\frac{e_{i,D}(n)}{\parallel \mathbf{u}_i(n) \parallel_2}\right)^2 + \frac{1}{2}\delta \parallel \mathbf{h}(n) \parallel_0$
IMSAF [12], [13]	$J(n) = \parallel \mathbf{h}(n+1) - \mathbf{h}(n) \parallel^2 + \sum_{i=1}^{N} \mathbf{\Lambda}_i [\mathbf{d}_{i,D}(n) - \mathbf{U}_i^T(n)\mathbf{h}(n+1)]$
IMSAF	$J(n) = \frac{1}{2}\sum_{i=1}^{N} \mathbf{e}_{i,D}^T(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n)$
BS-NSAF *	$J(n) = \frac{1}{2}\sum_{i=1}^{N} \frac{\lvert e_{i,D}(n)\rvert^2}{\parallel \mathbf{u}_i(n) \parallel^2} + \delta \parallel \mathbf{h}(n) \parallel_{2,0}$
BS-IMSAF *	$J(n) = \frac{1}{2}\sum_{i=1}^{N} \mathbf{e}_{i,D}^T(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n) + \delta \parallel \mathbf{h}(n) \parallel_{2,0}$

* Proposed in this paper.

What we propose in this paper can be summarized as follows:

- Establishment of the BS-NSAF. This algorithm has faster convergence speed than L_0-NSAF for BS system identification.
- Establishment of the BS-IMSAF. The BS-IMSAF has better convergence speed than BS-NSAF.
- Introducing the BS-DSR-IMSAF and BS-SR-IMSAF algorithms. These algorithms have lower computational complexity than BS-IMSAF.
- Studying the theoretical steady-state performance of proposed algorithms.
- Demonstrating of the proposed algorithms through several simulation results.

This paper is organized as follows: Sect. 2 describes the data model and IMSAF algorithm. In Sect. 3, the IMSAF is derived based on the gradient descent approach. In Sect. 4, the BS-NSAF algorithm is introduced. The family of BS-

IMSAF is proposed in 5. Sect. 6 studies the theoretical steady-state performance of the algorithms. The computational complexity of the proposed algorithm is discussed in Sect. 7. Finally, the paper ends with a comprehensive set of simulations supporting the validity of the results.

Throughout the paper, $(.)^{\mathrm{T}}$ represents transpose of a vector or matrix, $\|.\|_0$ indicates ℓ_0-norm of a vector, $\|.\|^2$ takes the squared Euclidean norm of a vector, $\|.\|_{2,0}$ creates $L_{2,0}$-norm of a vector, $[.]$ describes the Ceiling function, and $E\{.\}$ shows the Expectation.

Data Model and Review of IMSAF Algorithm

Consider a linear data model for the desired signal as

$$d(\mathrm{k})=\mathbf{u}^{\mathrm{T}}(\mathrm{k})\mathbf{h}^{\mathrm{o}}+\mathrm{v}(\mathrm{k}), \tag{1}$$

where $\mathbf{h}^{\mathrm{o}}$ is an unknown M-dimensional filter coefficients that we want to estimate, $v(k)$ is the additive noise with variance σ_v^2, and $\mathbf{u}(k)=[u(k),u(k-1),\dots,u(k-M+1)]^T$ denotes an M-dimensional input regressor vector. It is assumed that $v(k)$ is zero mean, white, Gaussian, and independent of $\mathbf{u}(k)$.

Fig. 1 shows the structure of the NSAF [7]. In this figure, $\mathbf{f}_1,\mathbf{f}_2,\dots,\mathbf{f}_N$ and $\mathbf{g}_1,\mathbf{g}_2,\dots,\mathbf{g}_N$, are analysis and synthesis filter impulse responses of N channel. The $u_i(k)$ and $d_i(k)$ are nondecimated subband signals. It is important to note that k refers to the index of the original sequences and n denotes the index of the decimated sequences (n=floor(k/N)). The decimated output signal is defined as $y_{i,D}(n)=\mathbf{u}_i^T(n)\mathbf{h}(n)$, $\mathbf{h}(n)=[h_1(n),h_2(n),\dots,h_M(n)]^T$ and $\mathbf{u}_i(n)=[u_i(nN),u_i(nN-1),\dots,u_i(nN-M+1)]^T$. Also, the decimated subband error signal is expressed as $e_{i,D}(n)=d_{i,D}(n)-\mathbf{u}_i^T(n)\mathbf{h}(n)$.

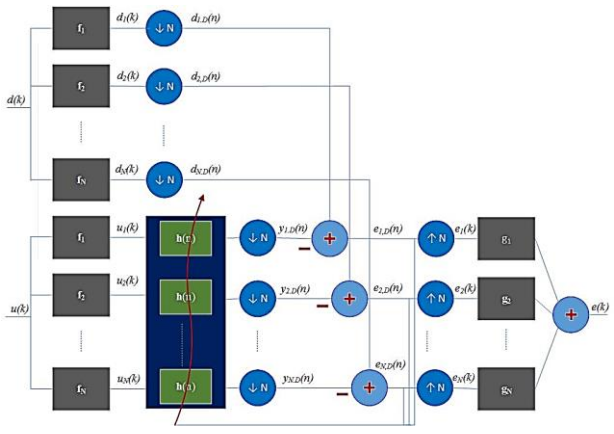


Fig. 1: Structure of the NSAF algorithm.

Now, by defining $[\mathbf{U}_i(n)]_{M\times K}$ and $[\mathbf{d}_{i,D}(n)]_{K\times 1}$ as

$$\mathbf{U}_i(n)=[\mathbf{u}_i(n),\mathbf{u}_i(n-1),\dots,\mathbf{u}_i(n-K+1)], \tag{2}$$

$$\mathbf{d}_{i,D}(n)=[d_{i,D}(n),\dots,d_{i,D}(n-K+1)]^T, \tag{3}$$

the IMSAF algorithm is derived from the solution of the following constraint optimization problem [12], [13],

$$\begin{aligned}J(n)=&\|\mathbf{h}(n+1)-\mathbf{h}(n)\|^2\\&+\textstyle\sum_{i=1}^{N}\boldsymbol{\Lambda}_i[\mathbf{d}_{i,D}(n)-\mathbf{U}_i^T(n)\mathbf{h}(n+1)],\end{aligned} \tag{4}$$

where $\boldsymbol{\Lambda}_i=[\lambda_{i,1},\lambda_{i,2},\dots,\lambda_{i,K}]$ is the Lagrange multipliers vector with length K. Using $\frac{\partial J(n)}{\partial \mathbf{h}(n+1)}=0$ and $\frac{\partial J(n)}{\partial \boldsymbol{\Lambda}_i}=0$, we get

$$\mathbf{h}(n+1)=\mathbf{h}(n)+\frac{1}{2}\sum_{i=1}^{N}\mathbf{U}_i(n)\boldsymbol{\Lambda}_i^T, \tag{5}$$

where

$$\boldsymbol{\Lambda}_i^T=2[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n), \tag{6}$$

and the output error vector, $[\mathbf{e}_{i,D}(n)]_{K\times 1}$, is given by

$$\mathbf{e}_{i,D}(n)=\mathbf{d}_{i,D}(n)-\mathbf{U}_i^T(n)\mathbf{h}(n), \tag{7}$$

Therefore, the update equation for IMSAF becomes

$$\mathbf{h}(n+1)=\mathbf{h}(n)+\mu\sum_{i=1}^{N}\mathbf{U}_i(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n), \tag{8}$$

where μ is the step-size.

Derivation of IMSAF Based on the Gradient Descent Method

In this section, we establish the IMSAF algorithm based on the gradient descent approach. Instead of minimizing (4), the following cost function is defined as [29].

$$J(n)=\frac{1}{2}\sum_{i=1}^{N}\mathbf{e}_{i,D}^T(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n),. \tag{9}$$

Based on the gradient descent approach, the filter coefficients recursion is given by

$$\mathbf{h}(n+1)=\mathbf{h}(n)-\mu\frac{\partial J(n)}{\partial \mathbf{h}(n)}, \tag{10}$$

Using (7) in (9), $J(n)$ becomes

$$\begin{aligned}J(n)=\frac{1}{2}\sum_{i=0}^{N-1}\{&\mathbf{d}_{i,D}^T(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{d}_{i,D}(n)\\&2\mathbf{h}^T(n)\mathbf{U}_i(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{d}_{i,D}(n)\\&\mathbf{h}^T(n)\mathbf{U}_i(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{U}_i^T(n)\mathbf{h}(n)\},\end{aligned} \tag{11}$$

and the gradient of $J(n)$ is

$$\begin{aligned}\frac{\partial J(n)}{\partial \mathbf{h}(n)}=&\sum_{i=1}^{N}\{-\mathbf{U}_i(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{d}_{i,D}(n)\\&+\mathbf{U}_i(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{U}_i^T(n)\mathbf{h}(n)\}\\=&-\textstyle\sum_{i=1}^{N}\{\mathbf{U}_i(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n)\},\end{aligned} \tag{12}$$

By substituting (12) in (10), the IMSAF algorithm is established as

$$\begin{aligned}&\mathbf{h}(n+1)\\&=\mathbf{h}(n)+\mu\sum_{i=1}^{N}\mathbf{U}_i(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n),\end{aligned} \tag{13}$$

Review of BS-LMS and derivation of BS-NSAF

In this section, we briefly review BS-LMS adaptive algorithms [27]. Then, the BS-NSAF is introduced.

A. Review of BS-LMS algorithm

In BS-LMS, a penalty of block-sparsity is inserted to the cost function of traditional LMS algorithms. This penalty is a mixed of $L_{2,0}$-norm of adaptive filter coefficients with equal group partition sizes. In BS-LMS, the cost function is defined as [27].

$$J(k) = |e(k)|^2 + \delta \parallel \mathbf{h}(k) \parallel_{2,0}, \tag{14}$$

where δ is positive factor to balance the estimation error and the penalty of block-sparsity. Also

$$\parallel \mathbf{h}(k) \parallel_{2,0} \approx \left\| \begin{bmatrix} \parallel \mathbf{h}_{[1]} \parallel_2 \\ \parallel \mathbf{h}_{[2]} \parallel_2 \\ \vdots \\ \parallel \mathbf{h}_{[B]} \parallel_2 \end{bmatrix} \right\|_0, \tag{15}$$

and

$$\mathbf{h}_{[i]} = [h_{(i-1)L+1}, h_{(i-1)L+2}, \cdots, h_{iL}]^T, \tag{16}$$

Denotes the ith block of $\mathbf{h}$. The parameters B and L are the number of blocks and the block partition size, respectively. Following the same strategy in L_0-LMS [18], the update equation for BS-LMS is given by

$$\mathbf{h}(k+1) = \mathbf{h}(k) + \mu e(k)\mathbf{u}(k) + \kappa \mathbf{f}(\mathbf{h}(k)), \tag{17}$$

where $\kappa = \mu\delta$ regulates the strength of block-sparse penalty for given step-size and zero attraction function

$$\begin{aligned} &\mathbf{f}(\mathbf{h}(k)) \\ &= [f_1(\mathbf{h}(k)), f_2(\mathbf{h}(k)), \cdots, f_M(\mathbf{h}(k))]^T, \end{aligned} \tag{18}$$

$$\begin{aligned} &f_j(\mathbf{h}(k)) \\ &= \begin{cases} \gamma^2 h_j(k) - \dfrac{\gamma h_j(k)}{\parallel \mathbf{h}_{[[j/L]]} \parallel_2}, \\ \quad when \quad 0 < \parallel \mathbf{h}_{[[j/L]]} \parallel_2 \leq \dfrac{1}{\gamma}, \\ 0, \quad otherwise. \end{cases} \end{aligned} \tag{19}$$

Fig. 2 shows zero attraction function for BS-LMS when $\gamma = 1$ and different length of blocks, $L = 1,2,4,8$.

If we have one block in BS-LMS, $L = 1$, then the zero attraction function in BS-LMS reduces to the zero attraction function in L_0-LMS. Zero attraction imposes an attraction to zero on small weight coefficients. After each iteration, a filter weight will decrease a little when it is positive, or increase a little when it is negative. Therefore, it seems that in space of weight coefficients, an attractor, which attracts the nonzero vectors, exists at the coordinate origin. The function of zero attractor improves the performance of LMS in sparse system identification. To be specific, in the adaptation process, a weight coefficient closer to zero shows a higher possibility of being zero itself in the impulse response.

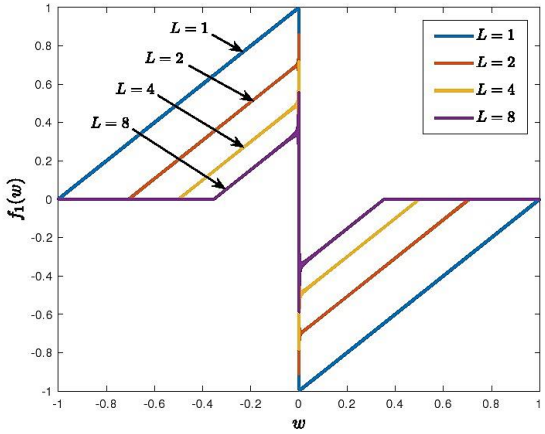


Fig. 2: Zero attraction function for BS-LMS when γ=1, L=1,2,4,8.

B. Proposed BS-NSAF algorithm

In [19], the L_0-NSAF was proposed. To improve the performance of NSAF and L_0-NSAF for block-sparse system identification, the BS-NSAF is presented. If the parameter K in (8) is set to 1, the NSAF algorithm is established. Therefore, by selecting this value in (9), the cost function for BS-NSAF algorithm is proposed as

$$J(n) = \frac{1}{2}\sum_{i=1}^{N} \frac{|e_{i,D}(n)|^2}{\parallel \mathbf{u}_i(n) \parallel^2} + \delta \parallel \mathbf{h}(n) \parallel_{2,0}, \tag{20}$$

Now, by applying the gradient descent approach in (10*J*) to the proposed cost function and setting $\frac{\partial J(n)}{\partial \mathbf{h}(n)}$ equal to zero, we get

$$\frac{\partial J(n)}{\partial \mathbf{h}(n)} = -\sum_{i=1}^{N} \frac{\mathbf{u}_i(n) e_{i,D}(n)}{\parallel \mathbf{u}_i(n) \parallel^2} + \delta \frac{\partial \parallel \mathbf{h}(n) \parallel_{2,0}}{\partial \mathbf{h}(n)}, \tag{21}$$

Finally, the update equation for BS-NSAF is established as

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \mu \sum_{i=1}^{N} \frac{\mathbf{u}_i(n) e_{i,D}(n)}{\parallel \mathbf{u}_i(n) \parallel^2} + \kappa \mathbf{f}(\mathbf{h}(n)), \tag{22}$$

where

$$\mathbf{f}(\mathbf{h}(n)) = [f_1(\mathbf{h}(n)), f_2(\mathbf{h}(n)), \cdots, f_M(\mathbf{h}(n))]^T, \tag{23}$$

The Family of BS-IMSAF Algorithms

Although, the IMSAF works well for dispersive unknown systems, its performance needs to be improved when the impulse response is block-sparse. In this section, three block-sparse adaptive algorithms are proposed. The first algorithm is BS-IMSAF algorithm. In the following, to reduce the computational complexity of BS-IMSAF algorithm, the selective regressors (SR) and dynamic selection of regressors (DSR) strategies are utilized and BS-SR-IMSAF and BS-DSR-IMSAF algorithms are derived.

A. The BS-IMSAF algorithm

By applying $L_{2,0}$-norm into (9), the optimization problem turns to

$$J(n) = \frac{1}{2}\sum_{i=1}^{N} \mathbf{e}_{i,D}^{T}(n)[\mathbf{U}_i^{T}(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n) + \delta \parallel \mathbf{h}(n) \parallel_{2,0}, \quad (24)$$

Setting $\frac{\partial J(n)}{\partial \mathbf{h}(n)}$ equal to zero,

$$\frac{\partial J(n)}{\partial \mathbf{h}(n)} = -\sum_{i=1}^{N} \{\mathbf{U}_i(n)[\mathbf{U}_i^{T}(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n)\} + \delta\frac{\partial \parallel \mathbf{h}(n) \parallel_{2,0}}{\partial \mathbf{h}(n)}, \quad (25)$$

the weight coefficients update equation becomes

$$\mathbf{h}(n+1) = \mathbf{h}(n) - \mu\delta\frac{\partial \parallel \mathbf{h}(n) \parallel_{2,0}}{\partial \mathbf{h}(n)} + \mu\sum_{i=1}^{N} \mathbf{U}_i(n)[\mathbf{U}_i^{T}(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n), \quad (26)$$

Finally, the weight update equation for BS-IMSAF is described as

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \kappa\mathbf{f}(\mathbf{h}(n)) + \mu\sum_{i=1}^{N} \mathbf{U}_i(n)[\mathbf{U}_i^{T}(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n), \quad (27)$$

where $\mathbf{f}(\mathbf{h}(n))$ is obtained from (24) and (25). Table 3 summarizes the BS-IMSAF algorithm.

Table 3: The BS-IMSAF algorithm

For $n = 0,1, \ldots$

$$\mathbf{u}(n) = [u(n),u(n-1), \ldots ,u(n-M+1)]^{T}$$

For $i = 1, \ldots ,N$

$$\mathbf{d}_{i,D}(n) = [d_{i,D}(n), \ldots ,d_{i,D}(n-K+1)]^{T}$$

$$\mathbf{U}_i(n) = [\mathbf{u}_i(n),\mathbf{u}_i(n-1), \ldots ,\mathbf{u}_i(n-K+1)]$$

$$\mathbf{e}_{i,D}(n) = \mathbf{d}_{i,D}(n) - \mathbf{U}_i^{T}(n)\mathbf{h}(n)$$

End

%——Update the filter:

$$\mathbf{f}(\mathbf{h}(n)) = [f_1(\mathbf{h}(n)),f_2(\mathbf{h}(n)), \cdots ,f_M(\mathbf{h}(n))]^{T}$$

$$f_j(\boldsymbol{h}(n)) = \begin{cases} \gamma^2 h_j(n) - \dfrac{\gamma h_j(n)}{\parallel \boldsymbol{h}_{[[j/L]]} \parallel_2}, \\ when \quad 0 < \parallel \boldsymbol{h}_{[[j/L]]} \parallel_2 \leq \dfrac{1}{\gamma}, \\ 0, \quad otherwise. \end{cases}$$

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \kappa\mathbf{f}(\mathbf{h}(n)) + \mu\sum_{i=1}^{N} \mathbf{U}_i(n)[\mathbf{U}_i^{T}(n)\mathbf{U}_i(n) + \epsilon\mathbf{I}]^{-1}\mathbf{e}_{i,D}(n)$$

End

B. *The BS-SR-IMSAF Algorithm*

In BS-SR-IMSAF, a subset of the input regressors at each subband is optimally selected for every adaptation. Let $\Theta_S = \{\theta_1,\theta_2, \ldots ,\theta_S\}$ denotes a S-subsets (subsets with S members) of the $\{0,1, \ldots ,K-1\}$. Now define

$$\mathbf{d}_{i,D,\Theta_S}(n) = [d_{i,D}(n-\theta_1), \ldots ,d_{i,D}(n-\theta_S)]^{T}, \quad (28)$$

and

$$\mathbf{U}_{i,\Theta_S}(n) = [\mathbf{u}_i(n-\theta_1), \ldots ,\mathbf{u}_i(n-\theta_S)], \quad (29)$$

Therefore, the output error vector is given by

$$\mathbf{e}_{i,D,\Theta_S}(n) = \mathbf{d}_{i,D,\Theta_S}(n) - \mathbf{U}_{i,\Theta_S}^{T}(n)\mathbf{h}(n), \quad (30)$$

The cost function for BS-SR-IMSAF is defined as

$$J_{\Theta_S}(n) = \delta \parallel \mathbf{h}(n) \parallel_{2,0} + \frac{1}{2}\sum_{i=1}^{N} \mathbf{e}_{i,D,\Theta_S}^{T}(n)[\mathbf{U}_{i,\Theta_S}^{T}(n)\mathbf{U}_{i,\Theta_S}(n)]^{-1}\mathbf{e}_{i,D,\Theta_S}(n). \quad (31)$$

Following the same approach in BS-IMSAF, we get

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \kappa\mathbf{f}(\mathbf{h}(n)) + \mu\sum_{i=1}^{N} \mathbf{U}_{i,\Theta_S}(n)[\mathbf{U}_{i,\Theta_S}^{T}(n)\mathbf{U}_{i,\Theta_S}(n)]^{-1}\mathbf{e}_{i,D,\Theta_S}(n), \quad (32)$$

We should select the regressors which makes $J_{\Theta_S}(n)$ as close as possible to $J(n)$. Thus, the optimum selection of the input regressors is obtained by a subset that minimizes

$$\Theta_S^{opt} = |\sum_{i=1}^{N} [\mathbf{e}_{i,D}^{T}(n)(\mathbf{U}_i^{T}(n)\mathbf{U}_i(n))^{-1}\mathbf{e}_{i,D}(n) - \mathbf{e}_{i,D,\Theta_S}^{T}(n)(\mathbf{U}_{i,\Theta_S}^{T}(n)\mathbf{U}_{i,\Theta_S}(n))^{-1}\mathbf{e}_{i,D,\Theta_S}(n)]|, \quad (33)$$

Since $\mathbf{e}_{i,D,\Theta_S}^{T}(n)(\mathbf{U}_{i,\Theta_S}^{T}(n)\mathbf{U}_{i,\Theta_S}(n))^{-1}\mathbf{e}_{i,D,\Theta_S}(n)$ is always smaller than $\mathbf{e}_{i,D}^{T}(n)(\mathbf{U}_i^{T}(n)\mathbf{U}_i(n))^{-1}\mathbf{e}_{i,D}(n)$, the optimum selection is reformulated by a subset that maximizes

$$\Theta_S^{opt} = \sum_{i=1}^{N} [\mathbf{e}_{i,D,\Theta_S}^{T}(n)(\mathbf{U}_{i,\Theta_S}^{T}(n)\mathbf{U}_{i,\Theta_S}(n))^{-1}\mathbf{e}_{i,D,\Theta_S}(n)], \quad (34)$$

To reduce the computational complexity of (36), we assume that the diagonal elements of $\mathbf{U}_{i,\Theta_S}^{T}(n)\mathbf{U}_{i,\Theta_S}(n)$ is much larger than off-diagonal elements [10], [12]. Therefore, (34) is approximated for each subband as

$$\mathbf{e}_{i,D,\Theta_S}^{T}(n)(\mathbf{U}_{i,\Theta_S}^{T}(n)\mathbf{U}_{i,\Theta_S}(n))^{-1}\mathbf{e}_{i,D,\Theta_S}(n) \approx \frac{e_{i,D}^{2}(n-\theta_1)}{\parallel \mathbf{u}_i(n-\theta_1) \parallel^2} + \cdots + \frac{e_{i,D}^{2}(n-\theta_S)}{\parallel \mathbf{u}_i(n-\theta_S) \parallel^2}, \quad (35)$$

where $\mathbf{e}_{i,D}(n) = [e_{i,D}(n),e_{i,D}(n-1), \ldots ,e_{i,D}(n-K+1)]^{T}$. Based on (35), the indices of the optimum subset at each subband for every iteration are obtained by the following simplified procedure:

1. Compute the following values for $0 \leq j \leq K-1$ and $1 \leq i \leq N$

$$\frac{e_{i,D}^{2}(n-j)}{\parallel \mathbf{u}_i(n-j) \parallel^2}, \quad (36)$$

2. The j-indices of Θ_S^{opt} for each i correspond to the indices of the S largest values of (36).

The BS-DSR-IMSAF Algorithm

In BS-DSR-IMSAF, the number of selected input regressors at each subband are dynamically changed for every adaptation. By defining the weight error vector as $\tilde{\mathbf{h}}(n) = \mathbf{h}^o - \mathbf{h}(n)$, the weight error vector update equation in BS-IMSAF can be stated as

$$\begin{aligned}\tilde{\mathbf{h}}(n+1) &= \tilde{\mathbf{h}}(n) - \kappa\mathbf{f}(\mathbf{h}(n)) \\ &\quad -\mu\textstyle\sum_{i=1}^{N}\mathbf{U}_i(n)[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}\mathbf{e}_{i,D}(n),\end{aligned} \tag{37}$$

Table 4 summarizes the BS-SR-IMSAF algorithm.

Table 4: The BS-SR-IMSAF algorithm

For $n = 0,1,\ldots$

$$\mathbf{u}(n) = [u(n),u(n-1),\ldots,u(n-M+1)]^T$$

For $i = 1,\ldots,N$

$$\mathbf{d}_{i,D}(n) = [d_{i,D}(n),\ldots,d_{i,D}(n-K+1)]^T$$

$$\mathbf{U}_i(n) = [\mathbf{u}_i(n),\mathbf{u}_i(n-1),\ldots,\mathbf{u}_i(n-K+1)]$$

$$\mathbf{e}_{i,D}(n) = \mathbf{d}_{i,D}(n) - \mathbf{U}_i^T(n)\mathbf{h}(n)$$

%——Determining the s-indices:

For $j = 0,1,\ldots,K-1$

compute $\frac{e_{i,D}^2(n-j)}{\|\mathbf{u}_i(n-j)\|^2}$

End

%——Update the desired signal vector:

$$\mathbf{U}_{i,\Theta_S}(n) = [\mathbf{u}_i(n-\theta_1),\ldots,\mathbf{u}_i(n-\theta_S)]$$

$$\mathbf{d}_{i,D,\Theta_S}(n) = [d_{i,D}(n-\theta_1),\ldots,d_{i,D}(n-\theta_S)]^T$$

$$\mathbf{e}_{i,D,\Theta_S}(n) = \mathbf{d}_{i,D,\Theta_S}(n) - \mathbf{U}_{i,\Theta_S}^T(n)\mathbf{h}(n)$$

End

%——Update the filter:

$$\mathbf{f}(\mathbf{h}(n)) = [f_1(\mathbf{h}(n)),f_2(\mathbf{h}(n)),\cdots,f_M(\mathbf{h}(n))]^T$$

$$f_j(\boldsymbol{h}(n)) = \begin{cases}\gamma^2 h_j(n) - \dfrac{\gamma h_j(n)}{\|\boldsymbol{h}_{[[j/L]]}\|_2}, \\ \quad when \quad 0 < \|\boldsymbol{h}_{[[j/L]]}\|_2 \le \dfrac{1}{\gamma}, \\ 0, \quad otherwise.\end{cases}$$

$$\begin{aligned}\mathbf{h}(n+1) &= \mathbf{h}(n) + \kappa\mathbf{f}(\mathbf{h}(n)) + \mu\sum_{i=1}^{N}\{\mathbf{U}_{i,\Theta_S}(n) \\ &\quad \times [\mathbf{U}_{i,\Theta_S}^T(n)\mathbf{U}_{i,\Theta_S}(n) + \epsilon\mathbf{I}]^{-1}\mathbf{e}_{i,D,\Theta_S}(n)\}\end{aligned}$$

End

Taking the squared Euclidean norm and then expectation from both sides of (39) leads to the mean-square deviation (MSD) that satisfies

$$E\,\|\,\tilde{\mathbf{h}}(n+1)\,\|^2 = E\,\|\,\tilde{\mathbf{h}}(n)\,\|^2 - \Delta, \tag{38}$$

where

$$\begin{aligned}\Delta = \sum_{i=1}^{N}&[\mu(2-\mu)E\{\mathbf{e}_{i,D}^T(n)(\mathbf{U}_i^T(n)\mathbf{U}_i(n))^{-1}\mathbf{e}_{i,D}(n)\} \\ &-2\mu\sigma_{v_{i,D}}^2\mathrm{Tr}(E[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1})] \\ &+\{\mathrm{crosstermswith}\kappa\},\end{aligned} \tag{39}$$

If Δ is maximized, then the fastest convergence is obtained. In (39), $\sigma_{v_{i,D}}^2$ is the variance of the ith subband signal of $v_i(n)$ being partitioned and decimated. We assume that the $\{\mathrm{crosstermof}\kappa\}$ are zero, because κ is very small value. Since the exact expected values are not available, the instantaneous values are used as follows

$$\begin{aligned}\hat{\Delta} = \mu(2-\mu)\sum_{i=1}^{N}&[\mathbf{e}_{i,D}^T(n)(\mathbf{U}_i^T(n)\mathbf{U}_i(n))^{-1}\mathbf{e}_{i,D}(n) \\ &-\tfrac{2}{2-\mu}\sigma_{v_{i,D}}^2\mathrm{Tr}[\mathbf{U}_i^T(n)\mathbf{U}_i(n)]^{-1}],\end{aligned} \tag{40}$$

Again we use the previous approximation for $\mathbf{U}_i^T(n)\mathbf{U}_i(n)$ and obtain [11], [12]

$$\begin{aligned}\hat{\Delta} = \mu(2-\mu)\sum_{i=1}^{N}&\{(\frac{e_{i,D}^2(n) - 2\sigma_{v_{i,D}}^2/(2-\mu)}{\|\,\mathbf{u}_i(n)\,\|^2}) \\ &+(\frac{e_{i,D}^2(n-1) - 2\sigma_{v_{i,D}}^2/(2-\mu)}{\|\mathbf{u}_i(n-1)\|^2}) + \cdots \\ &+(\frac{e_{i,D}^2(n-P+1) - 2\sigma_{v_{i,D}}^2/(2-\mu)}{\|\mathbf{u}_i(n-K+1)\|^2})\},\end{aligned} \tag{41}$$

From (41), we can find the following facts. If at each subband $e_{i,D}^2(n-j) > 2\sigma_{v_{i,D}}^2/(2-\mu)$, then $\mathbf{u}_i(n-j)$ contributes to maximizing $\hat{\Delta}$. However, if $e_{i,D}^2(n-j) \le 2\sigma_{v_{i,D}}^2/(2-\mu)$, then $\mathbf{u}_i(n-j)$ makes $\hat{\Delta}$ decrease. Therefore, we should perform the update with the input regressors satisfying $e_{i,D}^2(n-j) > 2\sigma_{v_{i,D}}^2/(2-\mu)$ at every iteration for the largest MSD decrease. Thus, the number of the selected input regressors at each subband for every iteration should be the same as the number of errors satisfying $e_{i,D}^2(n-j) > 2\sigma_{v_{i,D}}^2/(2-\mu)$.

Suppose $\Theta_{S_i(n)} = \{\theta_1,\theta_2,\ldots,\theta_{S_i(n)}\}$ indicates a subset with $S_i(n)$ members of the set $\{0,1.\ldots,K-1\}$ at each subband. Then, the update equation for proposed BS-DSR-IMSAF is introduced as

$$\begin{aligned}\mathbf{h}(n+1) &= \mathbf{h}(n) + \kappa\mathbf{f}(\mathbf{h}(n)) + \mu\sum_{i=1}^{N}\mathbf{U}_{i,\Theta_{S_i(n)}}(n) \\ &\quad [\epsilon\mathbf{I} + \mathbf{U}_{i,\Theta_{S_i(n)}}^T(n)\mathbf{U}_{i,\Theta_{S_i(n)}}(n)]^{-1}\mathbf{e}_{i,D,\Theta_{S_i(n)}}(n),\end{aligned} \tag{42}$$

where

$$\mathbf{d}_{i,D,\Theta_{S_i(n)}}(n) = [d_{i,D}(n-\theta_1),\ldots,d_{i,D}(n-\theta_{S_i(n)})]^T, \tag{43}$$

$$\mathbf{U}_{i,\Theta_{S_i(n)}}(n) = [\mathbf{u}_i(n-\theta_1),\ldots,\mathbf{u}_i(n-\theta_{S_i(n)})], \tag{44}$$

and

$$\mathbf{e}_{i,D,\Theta_{S_i(n)}}(n) = \mathbf{d}_{i,D,\Theta_{S_i(n)}}(n) - \mathbf{U}^T_{i,\Theta_{S_i(n)}}(n)\mathbf{h}(n). \quad (45)$$

The parameter $S_i(n)$ changes between 0 and K. The indices of the subset $(J_{S_i(n)})$ are obtained through the following procedure:

1. Compute the following values for $0 \le j \le K-1$ and $0 \le i \le N-1$

$$|e_{i,D}(n-j)| > \sqrt{\frac{2}{2-\mu}}\sigma_{v_{i,D}}, \quad (46)$$

2. The j-indices of $J_{S_i(n)}$ at each subband correspond to the indices that satisfies the condition in (46).

Table 5 summarizes the BS-DSR-IMSAF algorithm.

Table 5: The BS-DSR-IMSAF algorithm

For $n = 0,1, \dots$

$$\mathbf{u}(n) = [u(n), u(n-1), \dots, u(n-M+1)]^T$$

For $i = 1, \dots, N$

$$\mathbf{d}_{i,D}(n) = [d_{i,D}(n), \dots, d_{i,D}(n-K+1)]^T$$

$$\mathbf{U}_i(n) = [\mathbf{u}_i(n), \mathbf{u}_i(n-1), \dots, \mathbf{u}_i(n-K+1)]$$

$$\mathbf{e}_{i,D}(n) = \mathbf{d}_{i,D}(n) - \mathbf{U}_i^T(n)\mathbf{h}(n)$$

%——Determining the s-indices:

For $j = 0,1, \dots, K-1$

compute $|e_{i,D}(n-j)| > \sqrt{\frac{2}{2-\mu}}\sigma_{v_{i,D}}$

End

%——Update the desired signal vector:

$$\mathbf{U}_{i,\Theta_{S_i(n)}}(n) = [\mathbf{u}_i(n-\theta_1), \dots, \mathbf{u}_i(n-\theta_{S_i(n)})]$$

$$\mathbf{d}_{i,D,\Theta_{S_i(n)}}(n) = [d_{i,D}(n-\theta_1), \dots, d_{i,D}(n-\theta_{S_i(n)})]^T$$

$$\mathbf{e}_{i,D,\Theta_{S_i(n)}}(n) = \mathbf{d}_{i,D,\Theta_{S_i(n)}}(n) - \mathbf{U}^T_{i,\Theta_{S_i(n)}}(n)\mathbf{h}(n).$$

End

%——Update the filter:

$$\mathbf{f}(\mathbf{h}(n)) = [f_1(\mathbf{h}(n)), f_2(\mathbf{h}(n)), \cdots, f_M(\mathbf{h}(n))]^T$$

$$f_j(\boldsymbol{h}(n)) = \begin{cases} \gamma^2 h_j(n) - \dfrac{\gamma h_j(n)}{\| \boldsymbol{h}_{[[j/L]]} \|_2}, \\ when \quad 0 < \| \boldsymbol{h}_{[[j/L]]} \|_2 \le \dfrac{1}{\gamma}, \\ 0, \quad otherwise. \end{cases}$$

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \kappa\mathbf{f}(\mathbf{h}(n)) + \mu \sum_{i=1}^{N} \{\mathbf{U}_{i,\Theta_{S_i(n)}}(n) \times [\epsilon\mathbf{I} + \mathbf{U}^T_{i,\Theta_{S_i(n)}}(n)\mathbf{U}_{i,\Theta_{S_i(n)}}(n)]^{-1}\mathbf{e}_{i,D,\Theta_{S_i(n)}}(n)\}$$

End

Theoretical Performance Analysis

In this section, we analyze the performance of the family of BS-IMSAF algorithms. By defining the weight error vector, $\tilde{\mathbf{h}} = \mathbf{h}^o - \mathbf{h}(n)$, the general weight error vector update equation can be written as

$$\tilde{\mathbf{h}}(n+1) = \tilde{\mathbf{h}}(n) - \kappa\mathbf{f}(\mathbf{h}(n)) - \mu \sum_{i=0}^{N-1} \mathbb{U}_i(n)[\mathbb{U}_i^T(n)\mathbb{U}_i(n)]^{-1}\mathbb{e}_{i,D}(n), \quad (47)$$

where $\mathbb{U}_i(n)$ and $\mathbb{e}_{i,D}(n)$ is defined according to the Table 6.

Table 6: The definitions of $\mathbb{U}_i(n)$ and $\mathbb{e}_{i,D}(n)$

Algorithm	$\mathbb{U}_i(n)$	$\mathbb{e}_{i,D}(n)$
BS-IMSAF	$\mathbf{U}_i(n)$	$\mathbf{e}_{i,D}(n)$
BS-SR-IMSAF	$\mathbf{U}_{i,\Theta_S}(n)$	$\mathbf{e}_{i,D,\Theta_S}(n)$
BS-DSR-IMSAF	$\mathbf{U}_{i,\Theta_{S_i(n)}}(n)$	$\mathbf{e}_{i,D,\Theta_{S_i(n)}}(n)$

By taking the squared Euclidean norm from both sides of (25), we get

$$\| \tilde{\mathbf{h}}(n+1) \|^2 = \| \tilde{\mathbf{h}}(n) \|^2 , \quad (48)$$

$$-2\mu \sum_{i=0}^{N-1} [\tilde{\mathbf{h}}^T(n)\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) + 2\mu\kappa \sum_{i=0}^{N-1} [\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) + \mu^2 \sum_{i=0}^{N-1} \mathbb{e}_{i,D}^T(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) - \kappa\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n) + \kappa^2 \| \mathbf{f}(\mathbf{h}(n)) \|^2 , \quad (49)$$

From (1) and (7), we have

$$\tilde{\mathbf{h}}^T(n)\mathbb{U}_i(n) = \mathbb{e}_{i,D}^T(n) - \mathbb{v}_{i,D}^T(n), \quad (50)$$

Therefore, (49) is reformulated as

$$\| \tilde{\mathbf{h}}(n+1) \|^2 = \| \tilde{\mathbf{h}}(n) \|^2 - 2\mu \sum_{i=0}^{N-1} [\mathbb{e}_{i,D}^T(n) - \mathbb{v}_{i,D}^T(n)](\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) + 2\mu\kappa \sum_{i=0}^{N-1} \mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) + \mu^2 \sum_{i=0}^{N-1} \mathbb{e}_{i,D}^T(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) - \kappa\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n) + \kappa^2 \| \mathbf{f}(\mathbf{h}(n)) \|^2 , \quad (51)$$

which can be stated as

$$\| \tilde{\mathbf{h}}(n+1) \|^2 = \| \tilde{\mathbf{h}}(n) \|^2 - 2\mu \textstyle\sum_{i=0}^{N-1} \mathbb{e}_{i,D}^T(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) + 2\mu \textstyle\sum_{i=0}^{N-1} \mathbb{v}_{i,D}^T(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) + 2\mu\kappa \sum_{i=0}^{N-1} \mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbf{e}_{i,D}(n) + \mu^2 \textstyle\sum_{i=0}^{N-1} \mathbb{e}_{i,D}^T(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) - \kappa\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n) + \kappa^2 \| \mathbf{f}(\mathbf{h}(n)) \|^2 , \quad (52)$$

To simplify the recent relation, we apply the following independence assumptions [12], [30]:

1. $\mathbb{U}_i(n)$ is independent and identically distributed sequence matrix.

2. $\tilde{\mathbf{h}}(n)$ is independent of $\mathbb{U}_i(n)$

Now, taking the expectation from both sides of (52) and using the fact that $\mathbb{e}_{i,D}(n) = \mathbb{U}_i^T(n)\tilde{\mathbf{h}}(n) + \mathbb{v}_{i,D}(n)$, the third term in the right-hand side of (52) is simplified by

$$\sum_{i=0}^{N-1} E\{\mathbb{v}_{i,D}^T(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)\} = \sum_{i=0}^{N-1} E\{\mathbb{v}_{i,D}^T(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{v}_{i,D}(n)\} = \sum_{i=0}^{N-1}\sigma_{v_{i,D}}^2 \mathrm{Tr}(E[\mathbb{U}_i^T(n)\mathbb{U}_i(n)]^{-1}), \quad (53)$$

where $E\{\mathbb{v}_{i,D}(n)\mathbb{v}_{i,D}^T(n)\} = \sigma_{v_{i,D}}^2.\mathbf{I}$. Substituting (53) into (54), we obtain

$$\begin{aligned} E\parallel\tilde{\mathbf{h}}(n+1)\parallel^2 &= E\parallel\tilde{\mathbf{h}}(n)\parallel^2 \\ &-2\mu\sum_{i=0}^{N-1} E[\mathbb{e}_{i,D}^T(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)] \\ &+2\mu\sum_{i=0}^{N-1}\sigma_{v_{i,D}}^2\mathrm{Tr}(E[\mathbb{U}_i^T(n)\mathbb{U}_i(n)]^{-1}) \\ &+2\mu\kappa\sum_{i=0}^{N-1} E[\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)] \\ &+\mu^2\sum_{i=0}^{N-1} E[\mathbb{e}_{i,D}^T(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)] \\ &-\kappa E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)] + \kappa^2 E[\parallel\mathbf{f}(\mathbf{h}(n))\parallel^2], \end{aligned} \quad (54)$$

To simplify the recent relation, we assume that matrix, $\mathbb{U}_i^T(n)\mathbb{U}_i(n)$, is diagonal. This assumption was successfully applied in [10] and [11]. Using this assumption leads to

$$\begin{aligned} E\parallel\tilde{\mathbf{h}}(n+1)\parallel^2 &= E\parallel\tilde{\mathbf{h}}(n)\parallel^2 \\ &-\mu(2-\mu)\sum_{i=0}^{N-1} E\{(\frac{e_{i,D}^2(n)-2\sigma_{v_{i,D}}^2/(2-\mu)}{\parallel\mathbf{u}_i(n)\parallel^2}) \\ &+(\frac{e_{i,D}^2(n-1)-2\sigma_{v_{i,D}}^2/(2-\mu)}{\parallel\mathbf{u}_i(n-1)\parallel^2})+\cdots \\ &+(\frac{e_{i,D}^2(n-K+1)-2\sigma_{v_{i,D}}^2/(2-\mu)}{\parallel\mathbf{u}_i(n-K+1)\parallel^2})\} \\ &+2\mu\kappa\sum_{i=0}^{N-1} E[\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbf{U}_i(n))^{-1}\mathbf{e}_{i,D}(n)] \\ &-\kappa E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)] + \kappa^2 E[\parallel\mathbf{f}(\mathbf{h}(n))\parallel^2], \end{aligned} \quad (55)$$

The mean square deviation (MSD) and mean square error (MSE) are obtained by

$$MSD(n) = E\parallel\tilde{\mathbf{h}}(n)\parallel^2, \quad (56)$$

$$MSE(n) = E[e_{i,D}^2(n)], \quad (57)$$

Therefore,

$$\begin{aligned} MSD(n+1) &= MSD(n) \\ &-\mu(2-\mu)\sum_{i=0}^{N-1}(\frac{MSE(n)-2\sigma_{v_{i,D}}^2/(2-\mu)}{E\parallel\mathbf{u}_i(n)\parallel^2}) \\ &-\mu(2-\mu)\sum_{i=0}^{N-1}(\frac{MSE(n-1)-2\sigma_{v_{i,D}}^2/(2-\mu)}{E\parallel\mathbf{u}_i(n-1)\parallel^2})-\cdots \\ &-\mu(2-\mu)\sum_{i=0}^{N-1}(\frac{MSE(n-K+1)-2\sigma_{v_{i,D}}^2/(2-\mu)}{E\parallel\mathbf{u}_i(n-K+1)\parallel^2}) \\ &+2\mu\kappa\sum_{i=0}^{N-1} E[\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)] \\ &-\kappa E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)] + \kappa^2 E[\parallel\mathbf{f}(\mathbf{h}(n))\parallel^2], \end{aligned} \quad (58)$$

The recent relation can be rearranged as

$$\begin{aligned} MSD(n+1) &= MSD(n) \\ &-\mu(2-\mu)\sum_{i=0}^{N-1}(\frac{MSE(n)}{E\parallel\mathbf{u}_i(n)\parallel^2}+\cdots+\frac{MSE(n-K+1)}{E\parallel\mathbf{u}_i(n-K+1)\parallel^2}) \\ &+2\mu\sum_{i=0}^{N-1}\sigma_{v_{i,D}}^2(\frac{1}{E\parallel\mathbf{u}_i(n)\parallel^2}+\cdots+\frac{1}{E\parallel\mathbf{u}_i(n-K+1)\parallel^2}) \\ &+2\mu\kappa\sum_{i=0}^{N-1} E[\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)] \\ &-\kappa E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)] + \kappa^2 E[\parallel\mathbf{f}(\mathbf{h}(n))\parallel^2], \end{aligned} \quad (59)$$

when n goes to infinity, $MSD(n+1) = MSD(n)$, $MSE(n) = MSE(n-1) = \cdots = MSE(n-K+1)$, and $E\parallel\mathbf{u}_i(n)\parallel^2 = M\sigma_{u_i}^2$. Therefore, the above relation becomes

$$\begin{aligned} &\mu(2-\mu)\sum_{i=0}^{N-1}(\overbrace{\frac{MSE}{M\sigma_{u_i}^2}+\cdots+\frac{MSE}{M\sigma_{u_i}^2}}^{P=K,S,\mathrm{or}S_i(n)\mathrm{times}}) = \\ &2\mu\sum_{i=0}^{N-1}\sigma_{v_{i,D}}^2(\overbrace{\frac{1}{M\sigma_{u_i}^2}+\cdots+\frac{1}{M\sigma_{u_i}^2}}^{P=K,S,orS_i(n)\mathrm{times}}) \\ &+2\mu\kappa\sum_{i=0}^{N-1} E[\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbf{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)] \\ &-\kappa E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)] + \kappa^2 E[\parallel\mathbf{f}(\mathbf{h}(n))\parallel^2], \end{aligned} \quad (60)$$

Equation (60) can be simplified as

$$\frac{\mu(2-\mu)}{M}MSE\sum_{i=0}^{N-1}\frac{P}{\sigma_{u_i}^2} = \frac{2\mu}{M}\sum_{i=0}^{N-1}\frac{P\sigma_{v_{i,D}}^2}{\sigma_{u_i}^2} + \{\mathrm{terms with}\mathbf{f}(\mathbf{h}(n))\} \quad (61)$$

Finally, the steady-state MSE is given by

$$MSE = \frac{2\sum_{i=0}^{N-1}\frac{P\sigma_{v_{i,D}}^2}{\sigma_{u_i}^2}}{(2-\mu)\sum_{i=0}^{N-1}\frac{1}{\sigma_{u_i}^2}} + \frac{\{\mathrm{terms with}\mathbf{f}(\mathbf{h}(n))\}}{\frac{\mu(2-\mu)P}{M}\sum_{i=0}^{N-1}\frac{1}{\sigma_{u_i}^2}}, \quad (62)$$

The first term in the right-hand side of (62) is the steady-state MSE of IMSAF. The second term is related to L_0-IMSAF. Using $\kappa = \mu\delta$, we have

$$\overbrace{MSE}^{BS-IMSAF} = \overbrace{MSE}^{IMSAF} + \frac{2\mu\delta\sum_{i=0}^{N-1} E[\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)}{\frac{(2-\mu)P}{M}\sum_{i=0}^{N-1}\frac{1}{\sigma_{u_i}^2}} + \frac{-\delta E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)]+\mu\delta^2 E[\|\mathbf{f}(\mathbf{h}(n))\|^2]}{\frac{(2-\mu)P}{M}\sum_{i=0}^{N-1}\frac{1}{\sigma_{u_i}^2}} \quad (63)$$

By defining $a = \mu E\parallel\mathbf{f}(\mathbf{h}(n))\parallel^2$ and

$$b = 2\mu\sum_{i=0}^{N-1} E[\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n) - E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)], \quad (64)$$

the MSE relation becomes

$$\widetilde{MSE}^{BS-IMSAF} = \widetilde{MSE}^{IMSAF} + \frac{\delta^2 a + \delta b}{\frac{(2-\mu)P}{M}\sum_{i=0}^{N-1}\frac{1}{\sigma_{u_i}^2}}, \quad (65)$$

Since $0 < \mu < 2$, the denominator of second term in the right-hand side of (65) is positive, i.e. $\frac{(2-\mu)P}{M}\sum_{i=0}^{N-1}\frac{1}{\sigma_{u_i}^2} > 0$. Therefore, if $\delta^2 a + \delta b < 0$, then the MSE of family of BS-IMSAF algorithms will be lower than IMSAF. In the following, we find the condition for δ when $\delta^2 a + \delta b < 0$. The condition is $\delta < -b/a$. Based on this, we obtain

$$\delta < \frac{-2\mu\sum_{i=0}^{N-1} E[\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)]}{\mu E[\| \mathbf{f}(\mathbf{h}(n)) \|^2]} + \frac{E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)]}{\mu E[\|\mathbf{f}(\mathbf{h}(n))\|^2]}. \quad (66)$$

By using $\mathbb{e}_{i,D}(n) = \mathbb{U}_i^T(n)\tilde{\mathbf{h}}(n) + \nu_{i,D}(n)$ and independence assumptions, we have

$$E[\mathbf{f}^T(\mathbf{h}(n))\mathbb{U}_i(n)(\mathbb{U}_i^T(n)\mathbb{U}_i(n))^{-1}\mathbb{e}_{i,D}(n)] \approx E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)]. \quad (67)$$

Based on (69), the following condition is achieved

$$\delta < \frac{-2\mu\sum_{i=0}^{N-1} E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)] + E[\mathbf{f}^T(\mathbf{h}(n))\tilde{\mathbf{h}}(n)]}{\mu E[\| \mathbf{f}(\mathbf{h}(n)) \|^2]}, \quad (68)$$

Due to the analyze the relation in the steady-state, we need to replace the index n with ∞ in (68). By simplifying the recent relation, we obtain

$$\delta < \frac{(1-2\mu N)E[\mathbf{f}^T(\mathbf{h}(\infty))\tilde{\mathbf{h}}(\infty)]}{\mu E[\| \mathbf{f}(\mathbf{h}(\infty)) \|^2]}, \quad (69)$$

Finally, the steady-state MSE in the family of BS-IMSAF algorithms is given by

$$\widetilde{MSE}^{BS-IMSAF} = \widetilde{MSE}^{IMSAF} + \frac{(2\mu N - 1)\delta E[\mathbf{f}^T(\mathbf{h}(\infty))\tilde{\mathbf{h}}(\infty)] + \mu\delta^2 E[\| \mathbf{f}(\mathbf{h}(\infty)) \|^2]}{\frac{(2-\mu)P}{M}\sum_{i=0}^{N-1}\frac{1}{\sigma_{u_i}^2}}, \quad (70)$$

Computational Complexity

Table 7 presents the computational complexity of the IMSAF and the proposed algorithms in terms of the number of multiplications per iteration for real data. In this table, M is the filter length, N is the number of subbands, K is the number of input regressors, Q is the length of channel filters, L is the length of blocks, S is the number of selected input regressors, and $S_i(k)$ is the number of selected regressors at each subband which is dynamic. This table indicates that the number of multiplications in BS-IMSAF depends on K. But, in BS-SR-IMSAF and BS-DSR-IMSAF, this parameter depends on S and $S_i(k)$. Therefore, the computational complexity of BS-SR-IMSAF and BS-DSR-IMSAF is lower than BS-IMSAF. In the proposed algorithms, we have also additional $3M + LM$ multiplications and 1 division for the term of $\kappa\mathbf{f}(\mathbf{h}(n))$.

Table 7: The number of multiplications in IMSAF, BS-IMSAF, BS-SR-IMSAF, and BS-DSR-IMSAF algorithms

Algorithm	Number of Multiplications
IMSAF	$(K^2+2K)M + K^3 + K^2 + 3NQ$
BS-IMSAF^*	$(K^2+2K+3)M + K^3 + K^2 + 3NQ + LM$
BS-SR-IMSAF^*	$(S^2+2S+3)M + S^3 + S^2 + 2M(K-S) + 2K + 3NQ + LM$
BS-DSR-IMSAF^*	$\sum_{i=0}^{N-1}\frac{1}{N}[(S_i^2(k) + S_i(k) + K)M + S_i^3(k) + S_i^2(k)] + 3NQ + 3M + LM$

* Proposed in this paper.

Simulation Results

The performance of the proposed algorithms is evaluated by computer simulations in the system identification. To generate block sparse of impulse response, the Markov-Gaussian model is used as

$$P\{s_k = 0 | s_{k-1} = 0\} = p_1,$$
$$P\{s_k \neq 0 | s_{k-1} \neq 0\} = p_2.$$

where $p_1 = 0.99$, and $p_2 = 0.91$ [27]. The nonzero coefficients are generated according to the white Gaussian noise. The input signal is an AR(1) signal which is generated by passing a zero-mean white Gaussian noise through a first-order system $H(z) = \frac{1}{1-0.9z^{-1}}$. An additive white Gaussian noise was added to the system output, which sets the signal-to-noise ratio (SNR) to 40 dB. In all simulations, we show the normalized mean square deviation (NMSD), $10log_{10}(\frac{\|\mathbf{h}(n)-\mathbf{h}^o\|^2}{\|\mathbf{h}^o\|^2})$, which is evaluated by ensemble averaging over 200 independent trials. Table 8 shows the values of the parameters in the simulations. The impulse response of the unknown block sparse system with $M = 800$ has been presented in Fig. 3.

Fig. 4 shows the steady-state NMSD values versus κ for the family of BS-IMSAF algorithms. The values of κ changes from 10^{-8} to 10^{-3}. The optimum values for κ are observed in this simulation. Table 9 specifies the exact optimum values of κ. We observe that, the BS-NSAF and the family of BS-IMSAF algorithms have close optiumm values. Fig. 5 shows the NMSD learning curves in optimum values of κ for conventional and block sparse adaptive algorithms. We compared the learning curves of the proposed algorithms with NSAF [7], IMSAF [9], and L_0-NSAF [19] algorithms. The block sparse adaptive

algorithms show better convergence speed and lower steady-state error than classical NSAF and IMSAF algorithms. In comparison with BS adaptive algorithms, the NSAF and IMSAF algorithms have larger steady-state error. Also, the L_0-NSAF has larger steady-state error than BS adaptive algorithms.

Table 8: The values of the parameters in the simulations ($M = 800$, $S = 2$, and $\sigma_v^2 = 10^{-4}$).

Figure	N	K	L	μ	κ
— Performance for changing κ:					
Fig. 4	8	2	2	1	10^{-8},...,10^{-3}
Fig. 5	8	2	2	1	$\{1.7,3.8,8.5\} \times 10^{-6}$
Fig. 56	4	8	2	0.5	10^{-8},...,10^{-3}
Fig. 7	4	8	2	0.5	$\{0.36,1.74,2.59\} \times 10^{-6}$
— Performance of Changing L:					
Fig. 8	4	8	1,...,50	0.5	1.74×10^{-6}
— Performance for changing K:					
Fig. 10	4	2,...,8	5	0.5	5.5×10^{-6}
Fig. 11	4	2,...,8	5	0.5	5.5×10^{-6}
Fig. 12	4	2,...,8	5	0.5	5.5×10^{-6}

Table 9: Optimum values of κ in Figs. 4 and 6.

Algorithm	Optimum values of κ in Fig. 4	Optimum values of κ in Fig. 6
L_0-NSAF	1.7×10^{-6}	0.36×10^{-6}
BS-NSAF	8.5×10^{-6}	2.59×10^{-6}
BS-IMSAF	8.5×10^{-6}	2.59×10^{-6}
BS-SR-IMSAF	8.5×10^{-6}	2.59×10^{-6}
BS-DSR-IMSAF	3.8×10^{-6}	1.74×10^{-6}

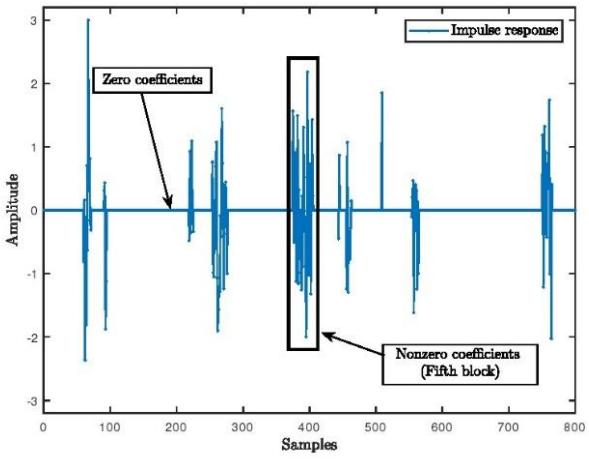


Fig. 3: The impulse response of unknown block sparse system.

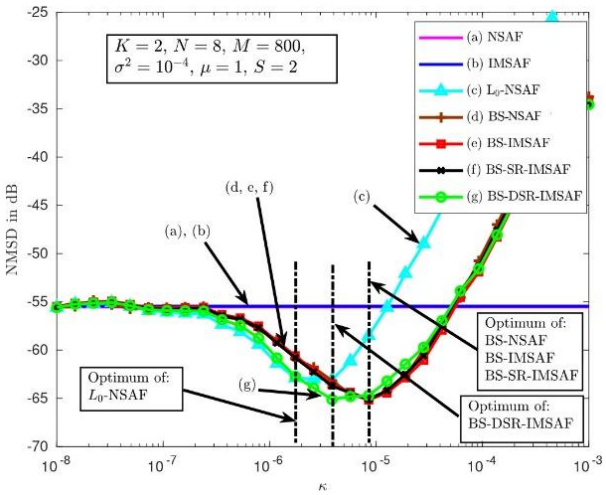


Fig. 4: The steady-state NMSD versus κ for the proposed algorithms ($M = 800$, $N = 8$, and $K = 2$).

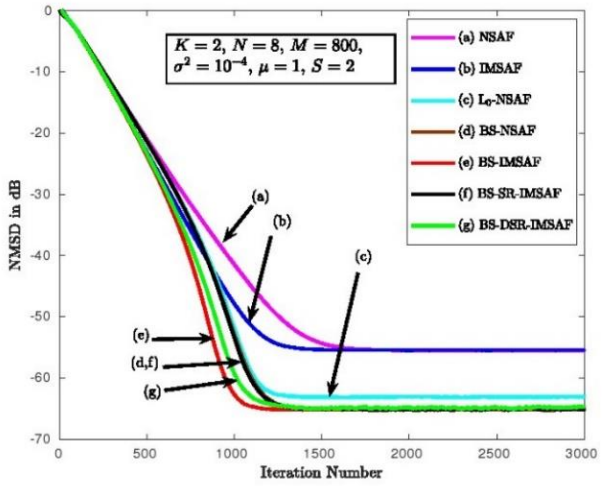


Fig. 5: The NMSD learning curves of all algorithms in the optimum values of κ (M = 800, N = 8, and K = 2).

Fig. 6 presents the steady-state NMSD values versus κ for the family of subband adaptive filter algorithms according to the parameters in Table 8.

Again, optimum values are obtained for block sparse adaptive filter algorithms. Table 9 shows the exact optimum values of κ in this simulation. We observe that BS-NSAF, BS-IMSAF, and BS-SR-NSAF have the same optimum values.

The BS-DSR-IMSAF has slightly lower optimum value than other BS algorithms. Fig. 7 compares the NMSD learning curves of classical and block sparse adaptive algorithms. This figure indicates that the proposed block sparse adaptive algorithms have better performance than NSAF, IMSAF, and L_0-NSAF algorithms. The performance of BS-SR-IMSAF and BS-DSR-IMSAF are close to the BS-IMSAF.

Furthermore, the steady-state NMSD of BS-SR-IMSAF and BS-DSR-IMSAF algorithms are lower than BS-IMSAF.

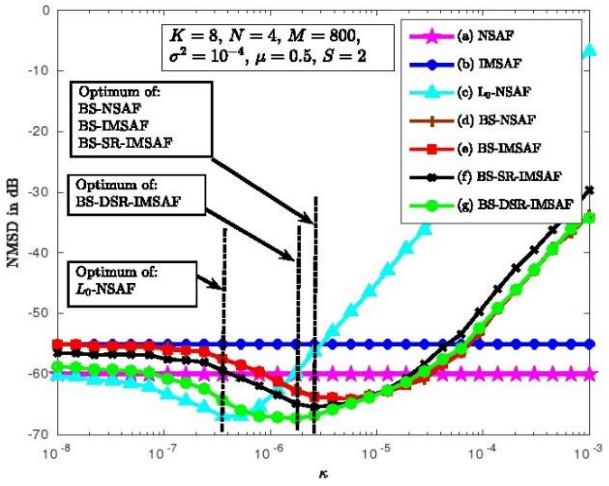


Fig. 6: The steady-state NMSD versus κ for the proposed algorithms (M = 800, N = 4, and K = 8).

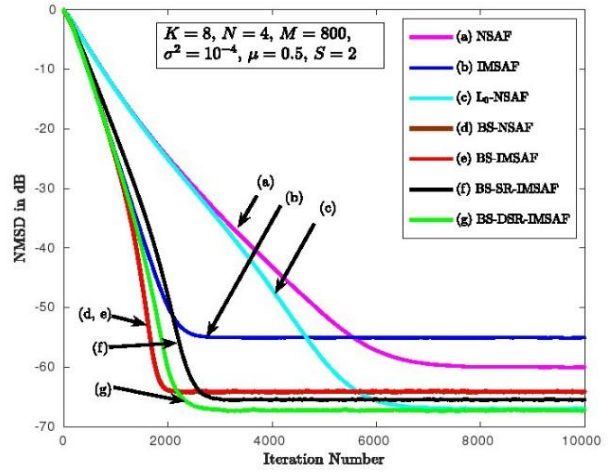


Fig. 7: The NMSD learning curves of all algorithms in the optimum values of κ (M = 800, N = 4, and K = 8).

In Fig. 8, we change the value of L and plot the steady-state NMSD values versus L. The values of L change from 1 to 50. The other parameters are set according to the Table 8. As we see, for L=5, the steady-state NMSD is obtained. Since the optimum value of κ changes between 10^{-5} and 10^{-6}, we select the midpoint value, 5.5×10^{-6}, for κ. Therefore, in the following simulations, we set the parameter L to 5 and the value of κ is set to 5.5×10^{-6}. Fig. 9 shows the performance of BS-IMSAF for different values of K. By increasing K, the convergence speed increases. But, the steady-state error also increases. The performance of BS-IMSAF is significantly better than other algorithms. Fig. 10 compares the learning curves of BS-SR-IMSAF and BS-IMSAF algorithm. We see that the BS-SR-IMSAF has the same performance as BS-IMSAF. But, the computational complexity of BS-SR-IMSAF is lower than BS-IMSAF. Fig. 11 investigates the performance of BS-DSR-IMSAF algorithm. Good performance can be seen for BS-DSR-IMSAF. By increasing the parameter K, the convergence speed of BS-DSR-IMSAF is faster than BS-IMSAF. Also, for BS-DSR-IMSAF, the same steady-state NMSD as BS-IMSAF is observed.

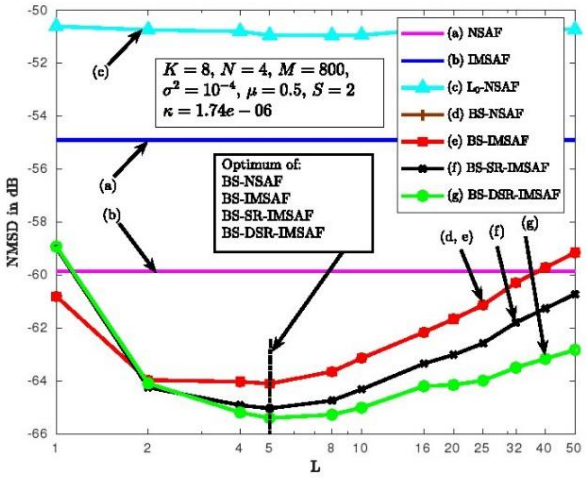


Fig. 8: The steady-state NMSD versus L, length of block, for NSAF, IMSAF and proposed BS-IMSAF algorithms ($M = 800$, $N = 4$, and $K = 8$).

Fig. 12 shows the number of selected regressors at each subband during the adaptation. This figure indicates that in the steady-state, the number of selected regressors converged to 1. It means that the steady-state error of this algorithm becomes low even for large values of K. Table 10 shows the simulated and theoretical steady-state NMSD for different values of SNR. These values are obtained for BS-IMSAF, BS-SR-IMSAF, and BS-DSR-IMSAF algorithms. As we see, the good agreement between simulated and theoretical steady-state NMSD values is observed. Table 11 compares the computation time and the values of NMSD in different algorithms at iterations 2000 and 3000. The parameters of the algorithms are according to the Fig. 9. This table indicates that the NMSD values of BS adaptive algorithms at iterations 2000 and 3000 are significantly lower than NSAF, IMSAF and L_0-NSAF algorithms.

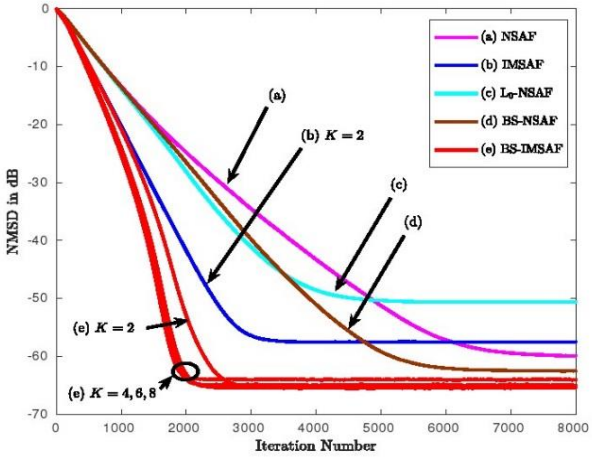


Fig. 9: The simulated and theoretical steady-state NMSD for different values of SNR (M = 800, N = 4, K = 8, μ = 0.5).

Table 10: The simulated and theoretical steady-state NMSD for different values of SNR ($\mathrm{M}=800$, $\mathrm{N}=4$, $\mathrm{K}=8$, $\mu=0.5$)

Algorithm	BS-IMSAF	BS-SR-IMSAF	BS-DSR-IMSAF
Simulation (SNR=10dB)	-14.45	-14.72	-14.85
Theory (SNR=10dB)	-15.21	-15.81	-15.90
Simulation (SNR=20dB)	-29.50	-29.60	-29.62
Theory (SNR=20dB)	-30.40	-30.35	-30.51
Simulation (SNR=40dB)	-64.81	-64.85	-64.89
Theory (SNR=40dB)	-65.20	-65.12	-65.20

Table 11: The computation time and the values of the NMSD in different algorithms at iterations 2000 and 3000

	Time (s)		NMSD in dB	
Algorithm	2000	3000	2000	3000
NSAF	2.2	3.3	-23.9	-34.5
IMSAF	56.4	84.6	-39.2	-57.2
L_0-NSAF	2.9	4.3	-26.3	-41.1
BS-NSAF	4.5	6.4	-25.1	-40.5
BS-IMSAF	56.9	85.5	-54.2	-64.8
BS-SR-IMSAF	14.5	21.7	-54.2	-64.8
BS-DSR-IMSAF	26.6	40.5	-63.5	-64.8

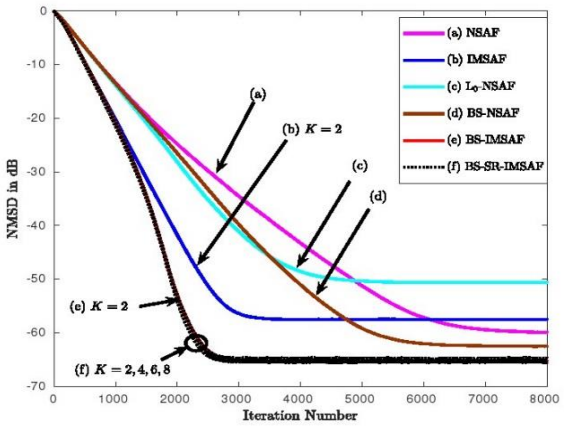


Fig. 10: The NMSD learning curves for different values of K, number of recent regressors in BS-SR-IMSAF algorithm ($\kappa = 5.5\times10^{-6}$,$\mathrm{M}=800$,$\mathrm{N}=4$,$\mathrm{K}=2,4,6,8$).

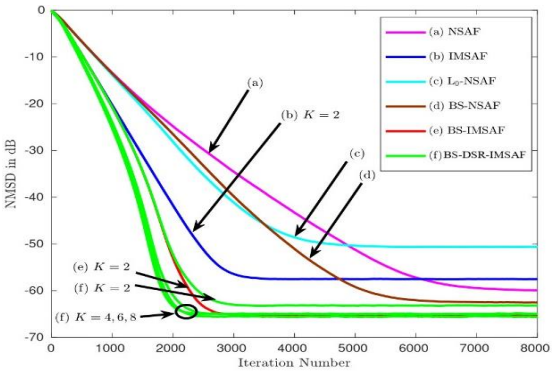


Fig. 11: The NMSD learning curves for different values of K, number of recent regressors in BS-DSR-IMSAF algorithm ($\kappa = 5.5\times10^{-6}$,$\mathrm{M}=800$,$\mathrm{N}=4$,$\mathrm{K}=2,4,6,8$).

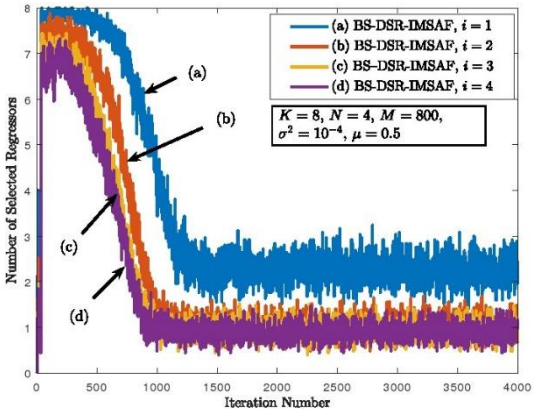


Fig. 12: The number of selected regressors at each subband in BS-DSR-IMSAF algorithm ($\mathrm{M}=800$, $\mathrm{N}=4$, and $\mathrm{K}=8$).

Summary and Conclusion

This paper presented the family of IMSAF algorithms for block sparse system identification. In the first algorithm, the BS-NSAF was introduced. This algorithm had better performance than NSAF for BS system identification. In the following the BS-IMSAF was presented. The proposed algorithm had better convergence speed than BS-NSAF. To reduce the computational complexity, the BS-SR-IMSAF and BS-DSR-IMSAF algorithms were developed. These algorithms had close performance to BS-IMSAF. Furthermore, the theoretical steady-state behavior of the proposed algorithms was studied.

Author Contributions

All authors contributed to all part of preparing and writing of this paper.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

LMS	Least Mean Squares
APA	Affine Projection Algorithm
NSAF	Normalized Subband Adaptive Filter
IMSAF	Improved Multiband Structured Subband Adaptive Filter
BS	Block Sparse
MSE	Mean Square Error
NMSD	Normalized Mean Square Deviation

References

[1] B. Widrow, S. D. Stearns, Adaptive Signal Processing, Englewood Cliffs, NJ: Prentice-Hall, 1985.

[2] J. R. Treichler, C. R. Johnson, M. G. Larimore, Theory and Design of Adaptive Filters, Wiley, 1987.

[3] S. Haykin, Adaptive Filter Theory, NJ: Prentice-Hall, 4th edition, 2002.

[4] A. H. Sayed, Adaptive Filters, Wiley, 2008.

[5] B. Farhang-Boroujeny, Adaptive Filters: Theory and Applications, Wiley, 1998.

[6] K. Ozeki, T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties, Electron. Commun. Jpn., 67-A: 19–27, 1984.

[7] K. A. Lee, W. S. Gan, "Improving convergence of the NLMS algorithm using constrained subband updates," IEEE Signal Process. Lett., 11 (9): 736–739, 2004.

[8] F. Yang, M. Wu, P. Ji, J. Yang, "An improved multiband-structured subband adaptive filter algorithm," IEEE Signal Process. Lett., 19: 647–650, 2012.

[9] F. Yang, M. Wu, P. Ji, J. Yang, "Low-complexity implementation of the improved multiband-structured subband adaptive filter algorithm," IEEE Trans. Signal Process., 63: 5133–5148, 2015.

[10] K. Y. Hwang, W. J. Song, "An affine projection adaptive filtering algorithm with selective regressors," IEEE Trans. Circuits Syst. II Express Briefs, 54(1): 43–46, 2007.

[11] S. J. Kong, K. Y. Hwang, W. J. Song, "An affine projection algorithm with dynamic selection of input vectors," IEEE Signal Process. Lett., 14(8): 529–532, 2007.

[12] M. S. E. Abadi, J. H. Husoy, M. J. Ahmadi, "Two improved multiband structured subband adaptive filter algorithms with reduced computational complexity," Signal Process., 154: 15–29, 2019.

[13] M. S. E. Abadi, M. J. Ahmadi, "Weighted improved multiband-structured subband adaptive filter algorithms," IEEE Trans. Circuits Syst. II Express Briefs, 2019.

[14] M. S. E. Abadi, M. J. Ahmadi, "Diffusion improved multiband-structured subband adaptive filter algorithm with dynamic selection of nodes over distributed networks," IEEE Trans. Circuits Syst. II Express Briefs, 66(3): 507–511, 2018.

[15] M. S. E. Abadi, H. Mesgarani, S. M. Khademiyan, "The wavelet transform-domain LMS adaptive filter employing dynamic selection of subband-coefficients," Digital Signal Process., 69: 94–105, 2017.

[16] D. L. Duttweiler, "Proportionate normalized least-meam-squares adaptation in echo cancellers," IEEE Trans. Speech Audio Process., 8(5): 508–518, 2000.

[17] A. Steingass, A. Lehner, F. Perez-Fontan, E. Kubista, B. Arbesser-Rastburg, "Characterization of the aeronautical satellite navigation channel through high-resolution measurement anf physical optics simulation," Int. J. Satell. Commun. Netw., 269: 1–305, 2008.

[18] Y. Gu, J. Jin, S. Mei, "l_0 norm constraint LMS algorithm for sparse system identification," IEEE Signal Process. Lett., 16(9): 774–777, 2009.

[19] Y. Yu, H. Zhao, B. Chen, "Sparse normalized subband adaptive filter algorithm with l_0-norm constraint," J. Franklin Inst., 353(18): 5121–5136, 2016.

[20] M. Lima, W. Martins, P. S. R. Diniz, "Affine projection algorithms for sparse system identification," in Proc. ICASSP: 5666–5670, 2013.

[21] M. Lima, T. Ferreira, W. Martins, P. S. R. Diniz, "Sparsity-aware data-selective adaptive filters," IEEE Trans. Signal Process., 62 (17): 4557–4572, 2014.

[22] L. Ji, J. NiK., "Sparsity-aware normalized subband adaptive filters with jointly optimized parameters," J. Franklin Instit., 357(17): 13144–13157, 2020.

[23] Y. Yu, T. Yang, H. Chen, R. Lamare, Y. Li, "Sparsity-aware SSAF algorithm with individual weighting factors: Performance analysis and improvements in acoustic echo cancellation," Signal Process., 178(1): 1–16, 2021.

[24] Y. Yu, H. Zho, R. Lamare, L. Lu, "Sparsity-aware subband adaptive algorithms with adjustable penalties," Digital Signal Process., 84(1): 93–106, 2019.

[25] Z. Habibi, H. Zayyani, M. S. E. Abadi, "A robust subband adaptive filter algorithm for sparse and block-sparse systems identification," J. Syst. Eng. Electron., 32(2): 487–497, 2021.

[26] E. Heydari, M. S. E. Abadi, S. M. Khademiyan, "Improved multiband structured subband adaptive filter algorithm with L_0-norm regularization for sparse system identification," Digital Signal Process., 122(4): 1–14, 2022.

[27] S. Jiang, Y. Gu, "Block-sparsity-induced adaptive filter for multi-clustering system identification," IEEE Trans. Signal Process., 63(20): 5318–5330, 2015.

[28] J. Liu, S. L. Grant, "Proportionate adaptive filtering for block-sparse system identification," IEEE/ACM Trans. Audio Speech Lang. Process., 24(4): 623–529, 2016.

[29] Z. Zhang, H. Zhao, "Affine projection M-estimate subband adaptive filters for robust adaptive filtering in impulsive noise," Signal Process., 120 (3): 64–70, 2016.

[30] H. C. Shin, A. H. Sayed, "Mean-Square performance of a family of affine projection algorithms," IEEE Trans. Signal Process., 52(1): 90–102, 2004.

Biographies

Esmail Heydari received the B.S. degree in Electrical Engineering from Hakim Sabzevari University in 2011, and the M.S. degree in Electrical Engineering from Shahid Rajaee Teacher Training University, Tehran, Iran, in 2017. Currently, he is a Ph.D. candidate at the Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. His research interests are adaptive filters and adaptive distributed networks.

- Email: eheydari@sru.ac.ir
- ORCID: 0009-0005-4223-4176
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mohammad Shams Esfand Abadi received the B.S. degree in Electrical Engineering from Mazandaran University, Mazandaran, Iran and the M.Sc. degree in Electrical Engineering from Tarbiat Modares University, Tehran, Iran in 2000 and 2002, respectively, and the Ph.D. degree in Biomedical Engineering from Tarbiat Modares University, Tehran, Iran in 2007. Since 2004 he has been with the Department of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran, where he is currently a professor. His research interests include digital filter theory, adaptive distributed networks, and adaptive signal processing algorithms.

- Email: mshams@sru.ac.ir
- ORCID: 0000-0002-9856-6592
- Web of Science Researcher ID: Y-7686-2019
- Scopus Author ID: 7006167272
- Homepage: https://www.sru.ac.ir/shams/

Seyed Mahmoud Khademiyan received the M.Sc. degree in Applied Mathematics from Iran University of Science and Technology, Tehran, Iran, in 2012. Currently, he is a Ph.D. candidate at the Department of Mathematics, Faculty of Science, Shahid Rajaee Teacher Training University, Tehran, Iran. His research interests include digital filter theory and adaptive signal processing algorithms.

- Email: m_khademiyan@sru.ac.ir
- ORCID: 0000-0002-9810-2819
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
E. Heydari, M. Shams Esfand Abadi, S. M. Khademiyan, "The new family of adaptive filter algorithms for block-sparse system identification," J. Electr. Comput. Eng. Innovations, 12(1): 133-146, 2024.

DOI: 10.22061/jecei.2023.10062.675

URL: https://jecei.sru.ac.ir/article_1988.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Multi-Task Learning Using Uncertainty for Realtime Multi-Person Pose Estimation

Z. Ghasemi-Naraghi, A. Nickabadi* *, R. Safabakhsh

Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran.

Article Info

Article History:
Received 24 June 2023
Reviewed 24 August 2023
Revised 12 September 2023
Accepted 08 October 2023

Keywords:
Realtime multi-person pose estimation
Multi-Task learning
Loss function
Task-dependent uncertainty

*Corresponding Author's Email Address: *nickabadi@aut.ac.ir*

Abstract

Background and Obejctives: Multi-task learning is a widespread mechanism to improve the learning of multiple objectives with a shared representation in one deep neural network. In multi-task learning, it is critical to determine how to combine the tasks loss functions. The straightforward way is to optimize the weighted linear sum of multiple objectives with equal weights. Despite some studies that have attempted to solve the realtime multi-person pose estimation problem from a 2D image, major challenges still remain unresolved.

Methods: The prevailing solutions are two-stream, learning two tasks simultaneously. They intrinsically use a multi-task learning approach for predicting the confidence maps of body parts and the part affinity fields to associate the parts to each other. They optimize the average of the two tasks loss functions, while the two tasks have different levels of difficulty and uncertainty. In this work, we overcome this problem by applying a multi-task objective that captures task-based uncertainties without any additional parameters. Since the estimated poses can be more certain, the proposed method is called "CertainPose".

Results: Experiments are carried out on the COCO keypoints data sets. The results show that capturing the task-dependent uncertainty makes the training procedure faster and causes some improvements in human pose estimation.

Conclusion: The highlight advantage of our method is improving the realtime multi-person pose estimation without increasing computational complexity.

Introduction

Multi-person pose estimation is an important open problem in computer vision. Human pose estimation (HPE) is widely used in many applications such as human-computer interaction, action recognition, motion capture, virtual reality, video surveillance, healthcare, gaming, and sports. HPE aims to automatically locate the human parts or keypoints (e.g. ankles, knees, hips, elbows) on images and videos. In many real-world applications, the desired HPE model is expected to: 1) run in realtime, 2) estimate the poses of several people simultaneously, and 3) extract poses from 2D images. Each one of these requirements introduces many challenges. The focus of this research is on the realtime localization of body parts of individuals in 2D images.

The challenging issues of the single-person pose estimation include the variety of clothes, scenes, body shapes, positions, and the scales of the persons in the scenes. The multi-person pose estimation imposes more challenges as an unknown number of people can appear in images at any position and scale. Interactions between people may cause occluded joints or interrupted limbs. In addition, if the runtime complexity of the solution grows with the number of people in the image, it may not be useful in some real-world applications.

The initial studies of the single-person pose estimation [1] were based on the pictorial structure models [2].

Traditionally, the focus was on hand-crafted features such as the histogram of oriented gradients (HOG). But, these methods have not shown promising generalization performance in detecting the accurate location of the body parts. Deep learning, especially the convolutional neural networks (CNNs), made a significant improvement in this field [3]. Some HPE approaches used famous deep neural networks such as ResNet [4] and Faster R-CNN [5] to detect keypoints more accurately [6], [7]. Another example of a DNN-based HPE model is a convolutional pose machine (CPM) which consists of a sequence of convolutional neural networks that repeatedly produce more precise 2D confidence maps for the locations of human body parts at each stage [8]. However, there is still a long way towards the complete resolution of dominating some challenges of single-person pose estimation such as occlusion of body parts and abnormal body poses.

Multi-person 2D pose estimation is a widely investigated form of HPE. The solutions provided for this task can be divided into two main categories: top-down and bottom-up. Top-down methods [7], [9]-[11] first detect the people in the image and then utilize single-person pose estimation for each individual. The speed and accuracy of top-down methods depend on the human detection speed and accuracy. Moreover, these models fail to estimate human poses in crowded scenes and nearby individuals. On the other hand, the bottom-up methods [12]-[16] for first, detect human body parts without knowing the number and location of people in the image; and after that, they associate the parts of each individual to each other. The inference time of the bottom-up methods is usually satisfactory and independent of the number of people while preserving high-quality results. However, these approaches suffer from difficulty in grouping body parts when there is a large overlap between people. Another weakness of most of the bottom-up methods is the low resolution of the position of the individuals, which can be solved by increasing the width of the network or defining an additional unit to compute the more precise locations for each candidate point. Considering the goals of this investigation, we follow the bottom-up methods.

Recent years have witnessed a huge growth in realtime multi-person 2D pose estimation research. The winner of the COCO Keypoints 2016 challenge, CMU-Pose, is the first realtime multi-person pose estimator on 2D images [14]. The newer version of this model, OpenPose [17], is an open-source library [18] to localize full-body points on single images. Several researchers have tried to improve the OpenPose method [16], [19], [20]. The prevailing methods are bottom-up methods that learn confidence maps and part affinity fields (PAFs) simultaneously. Confidence maps and PAFs locate body parts and limbs, respectively. Limb refers to the virtual line between two keypoints. PAFs are utilized in associating the detected body parts to each individual at the inference time. A confidence map is a gray-level image in which the pixel value refers to the likelihood of the intended part on it.

The above models intrinsically use a multi-task learning (MTL) approach in which the learning of confidence maps of the body parts and the PAFs can be treated as two different tasks. Most of them consider the average mean square error of the two outputs as the multi-task loss function. Although the two tasks have different levels of difficulty, they are given the same weight in the loss function. It has been shown that in MTL, finding appropriate weights of different tasks plays an important role [21]. In this work, we explain an MTL strategy for realtime multi-person pose estimation from a 2D image. The proposed model, called CertainPose, captures task-dependent uncertainty in a two-stream network that jointly produces confidence maps and PAFs. For the purpose of capturing uncertainty without increasing the parameters and computational complexity, the model is trained with a new loss function which is derived in this manuscript.

In summary, the main contribution of this research is twofold. First, a novel multi-task loss function is introduced that captures task-dependent uncertainty in multi regression tasks models. Second, a two-task architecture is trained by the new loss function for multi-person pose estimation. Our experiments show that the proposed model reduces the training time and improves the accuracy of the pose estimation without increasing the process time and trainable parameters.

This paper is organized as follows: First, the related literature is briefly reviewed in Section “Related Work”. Next, the proposed method is described. We report the results of the experiments in Section “Experiments”. Finally, the paper is concluded in Section “Conclusion”.

Related Work

Human pose estimation has been a popular subject of research in recent years. There are some invaluable surveys on HPE methods [22]-[25]. HPE problems are divided into single and multi-person human pose estimation problems. In this section, we summarize some of the most important 2D HPE methods and their cons and pros. We also review some studies which solve 3D pose estimation by incorporating depth information. Given that our innovation is focused on reducing uncertainty in HPE models, our next step is to review the existing literature on the sources and effects of uncertainty in pose estimation.

A. Single-Person Pose Estimation

One of the oldest methods for estimating and tracking the human pose is the motion capture technique in which

the performer has to wear markers (e.g. LED, magnetic, and reflective markers) near each joint so the joints can be easily identified. This has been a useful method in filmmaking and animation and is still useful for laboratory activities [26]. However, it requires special hardware and software to obtain and process data. In most real-world applications, it is necessary to estimate human pose without using markers.

The earlier approaches for human pose estimation from image or video consider a graphical structure to model the interactions between body parts obtained from local observations. The extracted features can be classified into low-level, mid-level, and high-level features with regard to human visual perception. Silhouette, contour, and edges are some famous low-level features. These features are not useful in situations with complex backgrounds and scenes. SIFT, Freak, and shapelet are known as mid-level features. HOG has been the most popular mid-level feature in HPE [25]. Context features [6], mixtures of parts [1] and PAFs [14] are examples of high-level features used for HPE. In addition to the aforementioned features extracted from the input image, body structure models are also employed in HPE to provide prior knowledge about the relation of different parts of the body. Kinematic models [2], cardboard models [27], and volumetric models [28] are the usual body structures used in the literature. Kinematic models consider a line for the connections between pairs of body parts and it is possible to define some priors about joint angles. The cardboard models are composed of information about body part rectangular shapes. Volumetric Models realistically represent 3D body shapes and poses.

The pictorial structure model (PSM) was the first model to recognize the objects based on the positions of their components. In PSM, objects are modeled with a graph in which nodes refer to the body components and edges refer to the relations of these components. Most PSM-based human pose estimation methods consider ten body parts as rectangles and find the best parameters of these rectangles (e.g. center, scale, and rotation) by using the extracted features and the angles between the pairs of body parts [1].

The emergence of deep neural networks significantly affected HPE as many other artificial intelligence applications. In 2014, the replacement of handcrafted features with the features extracted by convolutional neural networks made notable improvements in HPE [3]. As the first example, [3] optimizes an energy function which contains two parts: 1) a unary potential which identifies the body parts likelihood in all image pixels and 2) a pairwise potential that models the relations of neighbor parts by considering the relative location and the size of the angle between the links to the parent and child nodes.

Neural networks and probabilistic graphical models are two basic and useful tools in HPE that have exclusive weak points. In [29], both paradigms are combined to improve the HPE accuracy. This repetitive algorithm computes the likelihood of each part in all pixels of the image as a confidence map by using the prior of the intended part and the conditional likelihood of it given other parts.

To enable tractable inference, PSM-based methods have been restricted to tree-structured body models. Pose machine [30] is an iterative pose prediction algorithm that incorporates richer spatial interconnection among multiple parts and shares information across parts of different scales. The input of the pose machine model is an image that goes through multiple stages. Each stage includes multi predictors which predict confidence maps of different parts in different scales. Practically, feeding the output of the predictors of one stage to the next stage gradually improves confidence maps predictions.

The Convolutional Pose Machine brought about a significant improvement in single-person pose estimation accuracy and speed [8]. Actually, this model implements the pose machine idea by convolutional neural networks in multi-stages. Increasing the number of stages with a constant kernel size enlarges the receptive fields. Moreover, the multi stages of the algorithm improve the accuracy and confidence of estimating difficult parts' localizations by utilizing easy parts locations. In addition, the vanishing gradient problem of the deep neural networks is solved here by using intermediate supervision enforcing at the end of each stage.

The second winner of the COCO 2016 keypoints challenge [31] represents a method [6] based on ResNet [4]. First, they predict the confidence maps for body parts by a ResNet. The low resolution of the ResNet's outputs enforces estimating offsets for each part. This method is very accurate in predicting the pose, but due to the use of a very deep ResNet, it has a high computational complexity.

B. Multi-Person Pose Estimation

Multi-person pose estimation is more difficult than the single-person case due to the interactions between people, which increases the inference complexity. Increasing the number of people makes realtime performance a challenge for multi-person pose estimation models. Multi-person human pose estimation models can be divided into two main categories: top-down and bottom-up approaches. The top-down methods first detect each person in the image and then perform a single-pose estimation for each person. But, in the bottom-up methods, the human body parts in the image are first detected and then associated with each other to form humans and human poses. Although top-

down methods provide good accuracies, their speed and accuracy greatly depend on the human detection model. The computational cost of these models increases with the number of detected people. Also, crowded scenes and high interactions between people are challenging situations for top-down methods. In contrast, the bottom-up approach represents realtime methods with satisfying accuracy. The two challenges of the bottom-up methods are how to associate the parts to bodies and how to cope with the low resolution of each person in images that can be processed by the related neural networks. The latter problem can be resolved by increasing the width of the network or computing the precise locations of the body parts by searching the surrounding area of the approximate part locations. In this subsection, some top-down and bottom-up methods are described, respectively.

As the first example of the top-down methods, [10] proposes a probabilistic approach for parts grouping and labeling which uses HOG features for part detection. It is developed as a part-based approach by optimizing an articulated pictorial structure and a pixel-based method for image labeling. The multi-person human pose estimation is treated as an optimization problem with a single energy function. The goal of the inference step of this model is three-fold: 1) to determine the number of people and their locations, 2) to localize their joints, and 3) to assign every pixel of the input image to the background or a body part of a person.

A local joint-to-person method is presented for estimating the truncated or occluded poses in [11]. First, the people bounding boxes are detected by Faster R-CNN [5][5]. Then, the joint candidates are localized for each person and his neighbors by the convolutional pose machine [8]. In the end, a fully connected graph from the set of the detected joint candidates is constructed and the joint-to-person association is carried out locally with integer linear programming.

The second winner of COCO 2016 keypoint challenge [31] first detects the people in an image by a ResNet [4], and then, as described in the first part of this section, carries out the HPE by another ResNet [6]. Although this model provides accurate pose estimations, its computation complexity is high.

As the last top-down method, we refer to one of the state-of-the-art methods, Mask R-CNN [7]. Inspired by Faster R-CNN [5], Mask R-CNN is proposed which belongs to the top-down category of object detection models. The features are extracted using a standard convolutional neural network such as ResNet [4]. Some regions are suggested by the region proposal network (RPN) and then the proposed regions and extracted features aid to localize people and predict the confidence maps of each body part. The RPN and the body parts localization units have common feature extractor layers.

Deepcut [12] is one of the bottom-up multi-person pose estimation methods that performs the body part detection and pose estimation simultaneously. It employs an integer linear programming formulation to partition and label the set of body parts detected by a CNN-based part detector. It detects some candidates for body parts and determines their type, e.g. head, foot, and hand. Deepcut considers a complete graph on detected parts. Then, it solves the optimization problem by integer linear programming, for purpose of removing the edges and segmenting the graph into some disjointed subgraphs. As a result, each subgraph refers to a person's pose. Deepcut theory is satisfactory, but in practice, its speed is very slow. It needs about 72 hours for processing an image.

A deeper, stronger, and faster Deepcut method is proposed in Deepercut model [13] which uses a deeper part detector based on ResNet [4] and novel stronger image-conditioned pairwise terms in the objective function. Due to its pairwise and incremental optimization, Deepercut is faster than Deepcut. It first finds heads and shoulders locations. Then, elbows and wrists are added to the first stage solution and re-optimization is performed. Finally, the rest of the body parts are added to the previous stage solution and re-optimized. Yet, Deepercut is still too slow for realtime problems. It takes about 8 minutes for processing one image.

The winner of the COCO 2016 keypoints challenge [31] was the CMU-Pose method [14] which was motivated by the Convolutional Pose Machine [8]. The CMU-Pose method includes a feature extractor unit and multiple stages of convolutional neural networks. In each stage, confidence maps of each part and PAFs for encoding part-to-part associations are predicted and refined. PAFs are unit vectors defined for each pixel that show the direction of the limbs connecting body parts. The width of each limb is determined from the length of the connected line between the two parts. During test time, they compute the line integral over the corresponding PAFs along the lines connecting the candidate part locations.

The winner of PoseTrack 2017 challenge [15] improves the CMU-Pose method. They consider a deeper network for feature extraction and empirically increase the number of network stages from 6 to 7. The main contribution of this paper is the definition of enhanced PAFs. In the CMU-Pose method, $n-1$ PAFs are defined for the n body parts. But in this work, additional PAFs are considered. For example, in addition to PAFs between hip and knee and also between knee and ankle, they define additional PAFs between hip and ankle.

Another variant of the CMU-Pose model appeared in OpenPose which increases both speed and accuracy [17]. They released an open-source library which was the first

available system for realtime multi-person 2D pose estimation, including body, foot, hand, and facial keypoints [18]. They found that PAFs refinement is more important than confidence map refinement. So, they remove the part refinement stages and increase the depth of the network. An important aspect of OpenPose is that it includes the location of the feet in its pose estimation. Some applications such as filmmaking require foot information. In addition, the foot keypoints (e.g. big toe and heel) localization helps to estimate the whole-pose more accurately. To address these issues, a small subset of foot instances is labeled. OpenPose first obtains body and foot keypoints locations [14] and then runs hand and face keypoints detectors [32] for each detected person.

The deep Whole-body method [20] applies an MTL approach to the OpenPose network to train the model with different scale properties. To improve face and hand keypoints localization, the network increases the input resolution. Unfortunately, this implicitly reduces the effective receptive fields and therefore reduces body/foot localization accuracy. To solve this issue, the number of convolutional layers in each PAF stage is increased to recover the effective receptive field that was previously reduced. As the result, while the large receptive field is preserved, a high resolution for precise face and hand keypoint detection is provided. The new approach yields higher accuracy than that of the original OpenPose, especially for face and hand keypoint detection in occluded, blurry, and low-resolution images. Additionally, its total training time and inference runtime are less than the previous OpenPose.

While most of the HPE models improve the accuracy of the previous models by increasing the number of the model's parameters, Liu et al. propose a method for increasing the accuracy without very additional complexity [19]. Their contributions are resolution irrelevant encoding (RIE) and difficulty balanced loss (DBL). RIE is an inner block offset supervision that aids to learn the more precise locations for keypoints. Furthermore, DBL is a loss function containing two parts: 1) a Gaussian loss weight for different pixels which guides the network focus on useful information, and 2) the progressive punishment that discerns between left and right joints.

In [16], a lightweight architecture is designed to perform pose estimation on edge devices. They follow the OpenPose model, because of its quality and robustness to the number of people inside the frames. The parameters and complexity of the designed network are just 15 percent of the baseline 2-stage OpenPose with almost the same quality.

Pifpaf [33] is a multi-person human pose estimation method that is suited for low resolution and crowded scenes. They use two units, a part intensity field (PIF) to localize body parts and a part association field (PAF) to associate body parts with each other to form full human poses. Part Association Field predicts two vectors to the two parts at every image pixel. They use Laplace loss for regressions which incorporates a notion of uncertainty.

C. Pose Estimation by Using Depth

3D pose estimation is useful in widespread applications, such as human motion analysis, human-computer interaction, and robotics. A large number of approaches have been developed for pose estimation of one or several people, cars, or even dishes. When the depth information is available, 3D pose estimation is simple. However, it is possible to estimate depth from a monocular image or images from multiple camera views. As an example, [34] uses OpenPose with multiple synchronized video cameras for developing a 3D markerless motion capture technique. Here, we review some works which utilize or estimate depth to address their problem.

TesseTrack [35] is a top-down approach to estimate and track 3D body joints from a video in an end-to-end network. Central to this work is a novel spatio-temporal formulation that estimates a spatio-temporal volume around each person by a 4D CNN. The evaluation demonstrates the excellent performance of TesseTrack.

Occlusion-Net [36] is a self-supervised network that predicts 2D and 3D locations of occluded keypoints for objects, especially for cars. At the core of this network are two losses: 1) a trifocal tensor loss that provides indirect self-supervision for occluded keypoint locations that are visible in other views of the object, and 2) the self-supervised reprojection loss which estimates the 3D shape and camera pose.

In [37], the integration of bottom-up and top-down approaches is proposed to exploit their strengths. Their bottom-up network incorporates normalized heatmaps based on human detection, and their top-down network estimates human joints from all persons rather than from one. Finally, 3D poses are estimated from the top-down and bottom-up estimated 3D poses by an integration network. Also, to enforces natural two-person interactions, a two-person pose discriminator is proposed.

VoxelPose [38] estimates 3D poses of several persons from multiple camera views. It directly operates in the 3D space by aggregating the features in all camera views in the 3D voxel space. Then the features are fed into a network to localize all people. Finally, another network estimates a detailed 3D pose for each proposal.

A multi-stream multi-task network [39] for RGB-D-based human detection and head pose estimation is introduced to overcome challenges due to variations of illumination, clothing, resolution, pose, occlusion, and background. They integrate RGB, depth, and optical flow

data, as inputs to represent the appearance, shape, and motion information of humans, which makes full use of all the information provided by RGB-D video sequences to achieve state-of-the-art performance on three challenging datasets.

Depth sensors are prevalent in today's robotics, but large amount of data for training CNN is not available. Regarding the importance of object recognition and pose estimation from RGB-D images and the expensive cost of creating and annotating datasets for learning, [40] tries to address the problem with transfer learning. They propose a transfer learning from deep convolutional neural networks (CNN) that are pre-trained and provide a rich, semantically meaningful feature set. They transform depth data into a representation that is easily interpretable by a CNN trained on color images. Actually, instead of handcrafting or learning features, they relied on a convolutional neural network (CNN) which was trained on a large image dataset. They show that supervised learning on the CNN features outperforms state-of-the-art methods.

6D pose estimation is a type of pose estimation that is an important task in robotics. It is the task of detecting the 3D location and 3D orientation of an object. Given the depth information makes it feasible to extract the full 6D pose of object instances present in the scene. [41] uses analysis-by-synthesis which is a method to compare the observation with the generated output. They learn a CNN that compares observed and synthesized images. In particular, for pose estimation, a forward synthesis model generates images from possible poses and then selects the best match with the observed image.

D. Uncertainty in Pose Estimation

Despite the great achievements of deep neural networks in many applications, they still suffer from some weaknesses. While DNNs show excellent ability in perception, they fail in proper thinking and relational reasoning. DNNs are data-driven and need a lot of diverse data to learn a task perfectly. Practically, the insufficiency of the training data in terms of the number or diversity of the data increases the uncertainty of the DNNs' predictions. There are also uncertainties related to the nature of data and tasks. Capturing different kinds of uncertainties in training DNNs, especially in multi-task problems, may improve the efficiency of the training process and increase the accuracy of the developed model. In summary, three types of uncertainties are captured by Bayesian deep learning [42], [43]:

(1) Epistemic uncertainty is caused by the lack of data in training the deep model. If the test data is different from the training set, epistemic uncertainty increases more. Epistemic uncertainty can be resolved by increasing diverse training data or defining a prior distribution over the weights of the neural network. Some effective and simple algorithms are employed for estimating epistemic uncertainty [44]. For example, abnormal human poses which are not found in training data increase epistemic uncertainty in an HPE task.

(2) Heteroscedastic aleatoric uncertainty depends on the input data and differs from one to another input. Unlike epistemic uncertainty, heteroscedastic aleatoric uncertainty does not increase for out-of-date samples and does not decrease with more data. It is predicted by considering a distribution over the model outputs. As an example, the pose estimation of a person whose clothes' color or skin tone is very similar to the background is more uncertain than that of a person with distinct cloth color or skin tone. Modeling heteroscedastic uncertainties can be simple with less complexity.

(3) Homoscedastic aleatoric uncertainty does not depend on the inputs and is constant for all input data. Actually, it is related to tasks and hence is called task-dependent uncertainty. The common noises inherent in the observations or sensors cause this type of uncertainty in deep networks. For example, for estimation of human pose on a single 2D image, the lack of depth data causes uncertainty which is present in all outputs. It can be captured as the output of a model and can be decreased by utilizing other information e.g. estimated depth of an image. In this research, we only consider task-dependent uncertainty.

Uncertainty-Based Multi-Person Pose Estimation: CertainPose

Multi-task learning (MTL) is an efficient way to improve the learning of multiple tasks with a shared representation in a network. MTL increases the prediction accuracy by involving joint learning of various tasks. Besides, combining multiple objectives in a model reduces the computational complexity, so it is useful in realtime systems. A naive approach for learning multiple tasks is to minimize a weighted linear sum of multiple objectives with equal or fixed predefined weights. while it is important to determine the optimal weights [21], manual tuning of weights is difficult and inefficient.

Some studies are carried out to find an appropriate approach to combine the tasks' loss functions. As an example, the Cross-stitch network proposes a new unit that learns optimal coefficients for multiple objectives [45][45] In another attempt, Gradient Normalization (GradNorm), an adaptive multi-task loss balancing technique, normalizes across tasks instead of batch data in batch normalization [46]. The Human can learn from knowledge, but deep networks are data-driven and, unlike the probabilistic graphical models, cannot model the probability and the uncertainty well. Uncertainty is increased by the lack of training data and their diversity. Even if there is enough and diverse data, some uncertainties still remain due to the nature of the data

and the task.

Considering the different degrees of difficulty and uncertainty in various tasks, [47] learns the task-based uncertainties as additional parameters in the network and uses them to combine multi-task loss functions in an MTL loss function.

In recent years, the prevailing approach for multi-person pose estimation [15]-[17], [19], [20] has been the extraction of a set of shared features for learning two disjoint representations, simultaneously: confidence maps of the joints and part affinity fields (PAFs). Confidence maps show the likelihood of the presence of each keypoint at each pixel of the input image. On the other hand, PAFs represent limbs, the connections between keypoints, by a set of unit vectors. Actually, these models leverage MTL to learn two tasks with two distinct loss functions using a shared representation in a deep neural network.

In most networks which predict confidence maps and PAFs simultaneously, the simplest way of MTL is applied. The loss function of each task is the mean square error of the predicted and the actual outputs and the total loss function of the network is the average of the two objectives. However, the uncertainties of the two tasks are not necessarily equal, and using different weights for the two tasks boosts the learning efficiency. In this research, we intend to learn the two tasks more efficiently with the same computational complexity as the base networks. The main novelty of this work is to capture task-dependent uncertainties in an MTL method without any additional parameters. So, we call the proposed method "CertainPose", as it captures the uncertainty and estimates more certain poses. This section continues by describing the overall architecture of our model. Then, we explain three types of uncertainties and an MTL approach that model task-dependent uncertainties. Finally, a new loss function is introduced for training confidence maps and PAFs more fairly, which captures task-dependent uncertainty without any additional parameters.

A. Network Architecture

The role of the CertainPose model is to predict confidence maps and PAFs for the input image, as shown Fig. 1. There are three main components: 1) Feature extractor, 2) Confidence maps predictor, and 3) PAFs predictor. The input of the model is an image that is first fed into a feature extractor, and then the extracted features are used by the Confidence maps predictor and PAFs predictor. The second component predicts the confidence map of each body joint which represents the confidence score of that joint at the location at each pixel. The third component predicts the PAFs for each body limb, consisting of a directional vector at each pixel of the input image. While the input and architecture of these two networks are similar, their goals are different. They are trained in parallel, but with different loss functions and ground truth maps and fields, which result are yielded from the ground truth coordinates of multi-person poses. Therefore, the parameters of the two networks are trained with different loss values. To calculate an overall loss value for training the CertainPose network, the loss functions of the two parallel networks can be aggregated in different ways. We propose a new loss function that is described later.

The detailed architecture of the CertainPose model is outlined in Fig. 2. First, the input image (I) is fed into the feature extraction unit. The extracted features are then passed to two branches of the model, each of which consisting of a series of convolutional and pooling layers. In the first branch, a feedforward network predicts a set of 2D confidence maps (S) for the body parts' locations. There are J confidence maps for different parts (keypoints) of the body in $S = \{S_1, S_2, \ldots, S_J\}$. In the second branch, a set of 2D vector fields (L) of PAFs are predicted which encode the unit vectors in the direction of limbs, resembling the connections between adjacent body parts. The PAFs set consists of C PAF related to the C limbs $L = \{L_1, L_2, \ldots, L_C\}$. The two-branch network is repeated over t successive stages to refine the predictions. At each stage, the confidence maps and PAFs of the previous step along with the extracted features are taken as the input and the refined confidence maps and PAFs are generated as the outputs.

The internal structure of the feature extractor and the units of the two branches of the network are demonstrated in Fig. 2 Similar to CMU-Pose [17], CertainPose uses the first ten layers of the VGG-19 network as the feature extractor, and adds two more 3x3 convolutional layers to these layers.

The two branches of the network consist of 6 stages. In the first stage, shared features are fed into two disjoint CNN networks with the same layers: two 3x3 convolutional layers followed by two 1x1 convolutional layers regressing the tasks' outputs, i.e. the confidence maps and PAFs.

The inputs of the next successive stages consist of the concatenation of the shared features and the outputs of the previous stage (confidence maps and PAFs).

In stages 2 to 6, there are two CNN networks in each stage which are similar to each other with four 7x7 convolutional layers and two 1x1 convolutional layers. The number of parameters in each layer of CertainPose is shown in Fig. 3.

Moreover, ReLu is used as the activation function in all neurons.

The deep networks suffer from the vanishing gradient problem.

The intermediate supervision at each stage addresses this problem by replenishing the gradient periodically.

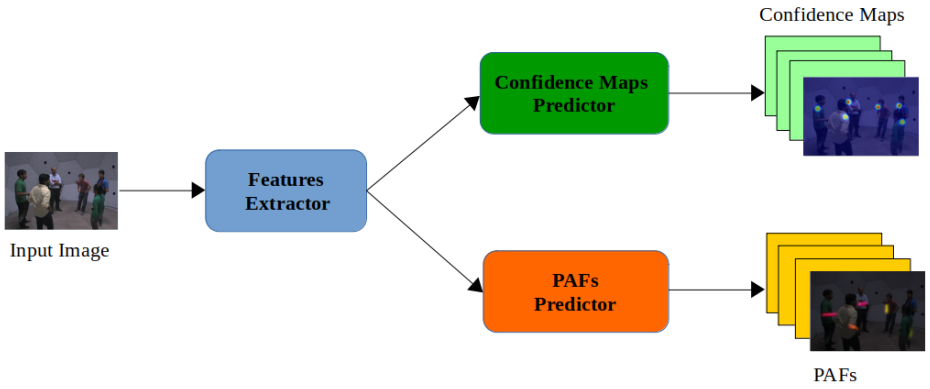


Fig. 1: The block diagram of CertainPose model.

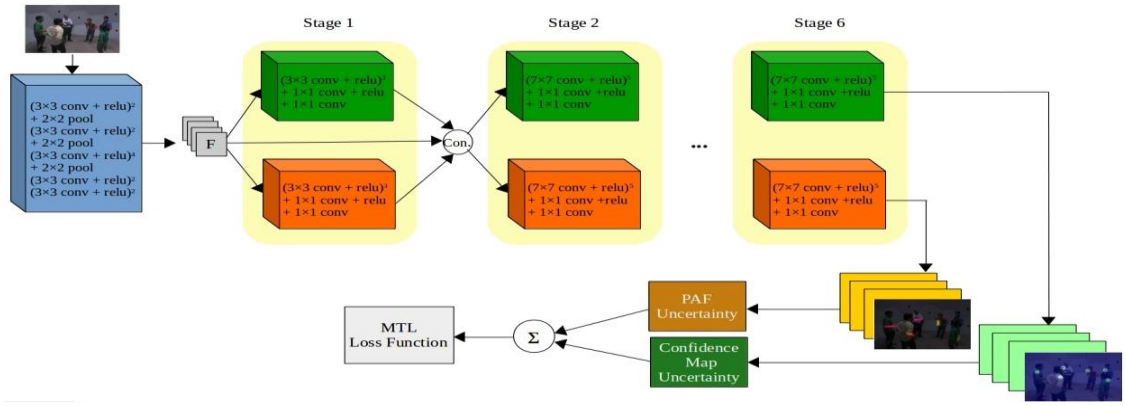


Fig. 2: The detailed architecture of CertainPose.

Number Of Parameters				
Feature Extractor layers	Confidence maps predictor		PAFs predictor	
1792	Stage 1	Stage 2 to 6	Stage 1	Stage 2 to 6
36928	147584	1160448	147584	1160448
73856	147584	802944	147584	802944
147584	147584	802944	147584	802944
295168	66048	802944	66048	802944
590080	19494	802944	9747	802944
590080		16512		16512
590080		4902		2451
1180160				
2359808				
1179904				
295040				
Total Parameters			**52,311,446**	
Trainable Parameters			**44,970,966**	
Non-Trainable Parameters			**7,340,480**	

Fig. 3: The number of parameters of CertainPose layers.

To guide the network to estimate the confidence maps and PAFs more accurately, the network is trained with a multi-task loss function. The loss function of each task is the mean of squared differences between the estimated and the actual outputs.

As the total loss function, CMU-Pose and the other approaches following it consider the average of the two tasks' loss functions. But, we attempt to have a fairer loss function for learning the two objectives by considering task-dependent uncertainties for the two tasks. Because of the more certain estimated poses, the new model is called 'CertainPose'. The new loss function is derived in subsection "Loss Function".

At the inference step, CertainPose predicts PAFs and Confidence maps for the input image. Similar to [14], non-maximum suppression is carried out to discretize the confidence maps and obtain some candidates for each part. A graph is then formed using candidate parts as vertices and candidate limbs as edges. To perform multi-person pose estimation, we should parse the graph and select the optimal set of limbs by measuring the association scores of the edges and removing the non-optimal edges. The score of a candidate limb is calculated by the line integral over the corresponding PAF along the candidate limb, virtual line segment connecting the candidate parts.

To speed up the parsing procedure, we use a greedy method. A greedy algorithm is a problem-solving approach that involves selecting the most advantageous

option at each step. While this strategy may not yield the best solution in all cases, it can produce locally optimal solutions that approximate the global optimal solution. Although greedy algorithms are not guaranteed to find the best solution, they are known for their speed and simplicity, making them a popular choice in real-time applications. We first consider a spanning tree skeleton for the human body instead of a complete graph, e.g. we ignore the virtual connected line between the head and the elbow. Then, we solve a bipartite matching problem to detect each limb. Bipartite matching is finding a set of edges between two vertices of two disjoint sets of vertices in the way that no two edges share an endpoint. For example, if we have three head and three shoulder keypoints, we should find the best three edges between the heads and shoulders without any shared point. Bipartite matching of disjoint parts' pairs obtains the limb connection candidates for each limb independently. Therefore, we can estimate the full-body poses of multiple people by assembling the candidate limbs.

B. Uncertainty

As described before, we only consider task-dependent uncertainty. Due to the importance of capturing task-based uncertainties and appropriately weighting the losses in multi-task learning, the uncertainty-based weighting method seems to be better than equal weighting of the losses [47]. The weights can be learned as a part of the convolutional neural network and loss functions. If the probabilistic likelihood of a regression task is considered as a Gaussian distribution, the variance parameter represents the noise and uncertainty of the task. In the following subsection, we describe this approach.

The problem (1) is finding the best weights w for the multi-task network f using the training data set $\{(I^{(i)}, S^{(i)}, L^{i}): i = 1,2, \dots, N\}$ where $I^{(i)}$, $S^{(i)}$, and $L^{(i)}$ refer to i-th sample input image, output set of confidence maps and PAFs, respectively.

$$\arg \max_{w} \mathcal{J}(w) \tag{1}$$

where

$$\mathcal{J}(w) = \prod_{i=1}^{N} p\left(S^{(i)}, L^{(i)} | I^{(i)}, w\right)$$
$$S = \mathcal{N}(f^{w_S}(I), \sigma_S^2); \; w_S \subset w$$
$$L = \mathcal{N}(f^{w_L}(I), \sigma_L^2); \; w_L \subset w \tag{2}$$

where w_S and w_L are weights related to the confidence maps and PAFs regression tasks, respectively. We assume that the two tasks are independent, so $\mathcal{J}(w)$ can be written as (3).

$$\mathcal{J}(w) = \prod_{i=1}^{N} p\left(S^{(i)} | I^{(i)}, w_S\right) p\left(L^{(i)} | I^{(i)}, w_L\right)$$
$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_S} exp\left(\frac{-\| S^{(i)} - f^{w_S}(I^{(i)}) \|^2}{2\sigma_S^2}\right)$$
$$\frac{1}{\sqrt{2\pi}\sigma_L} exp\left(\frac{-\| L^{(i)} - f^{w_L}(I^{(i)}) \|^2}{2\sigma_L^2}\right) \tag{1}$$

We solve the problem by minimizing the loss function, $\mathcal{L}$, instead of maximizing $\mathcal{J}$ (4).

$$\max_{w} \mathcal{J}(w) \equiv \min_{w} \mathcal{L}(w)$$
$$\mathcal{L}(w) = -\log\left(\mathcal{J}(w)\right)$$
$$= \sum_{i=1}^{N} \log\left(\sqrt{2\pi}\sigma_S\right) + \frac{\| S^{(i)} - f^{w_S}(I^{(i)}) \|^2}{2\sigma_S^2}$$
$$+\log\left(\sqrt{2\pi}\sigma_L\right) + \frac{\| L^{(i)} - f^{w_L}(I^{(i)}) \|^2}{2\sigma_L^2}$$
$$\mathcal{L}(w) = N\log(\sigma_S) + N\log(\sigma_L)$$
$$+\frac{1}{2\sigma_S^2}\sum_{i=1}^{N} \| S^{(i)} - f^{w_S}(I^{(i)}) \|^2$$
$$+\frac{1}{2\sigma_L^2}\sum_{i=1}^{N} \| L^{(i)} - f^{w_L}(I^{(i)}) \|^2 \tag{3}$$

$$\mathcal{L}(w) = \log(\sigma_S) + \log(\sigma_L)$$
$$+\frac{1}{2\sigma_S^2}\mathcal{L}_S(w_S) + \frac{1}{2\sigma_L^2}\mathcal{L}_L(w_L)$$
$$\mathcal{L}_S(w_S) = \frac{1}{N}\sum_{i=1}^{N} \| S^{(i)} - f^{w_S}(I^{(i)}) \|^2$$
$$\mathcal{L}_L(w_L) = \frac{1}{N}\sum_{i=1}^{N} \| L^{(i)} - f^{w_L}(I^{(i)}) \|^2 \tag{4}$$

Equation (5) shows the final solution of the proposed loss function, where σ_S and σ_L should be learned. As shown in this section, the task-dependent uncertainty can be captured by multi-task learning, while learning the additional parameters increases computational complexity.

C. Loss Function

We propose a new method for capturing the task-based uncertainties. In this method, the network architecture does not change and no further computational complexity is required. The new loss function (6) is derived as (7).

$$\arg \min_{w} \mathcal{L}(w) \tag{5}$$

where

$$\mathcal{L}(w) = \sum_{i=1}^{N} \log(\sigma_S) + \log(\sigma_L) + \frac{1}{2\sigma_S^2}\mathcal{L}_S(w_S) + \frac{1}{2\sigma_L^2}\mathcal{L}_L(w_L)$$
$$\mathcal{L}_S(w_S) = \frac{1}{N}\sum_{i=1}^{N} \| S^{(i)} - f^{w_S}(I^{(i)}) \|^2$$
$$\mathcal{L}_L(w_L) = \frac{1}{N}\sum_{i=1}^{N} \| L^{(i)} - f^{w_L}(I^{(i)}) \|^2 \quad (6)$$

In regression tasks, the likelihood is considered as a Gaussian whose mean is the output of the model (8). This means when the network's parameters are learned through the training process, the network's output for each task approaches the mean of the corresponding response variable given the input, i.e. $\overline{S}$ and $\overline{L}$ for our two tasks. So, σ_S^2 and σ_L^2 can be estimated with sample variances $\overline{Var_S^2}$ and $\overline{Var_L^2}$, respectively.

$$p(S_i \mid \overline{S}) = \mathcal{N}(\overline{S}, \sigma_S^2)$$
$$p(L_i \mid \overline{L}) = \mathcal{N}(\overline{L}, \sigma_L^2) \quad (7)$$

$$\mathcal{L}_S(w_S) = \frac{1}{N}\sum_{i=1}^{N} \| S^{(i)} - \overline{S} \|^2 = \overline{Var_S^2}$$
$$\mathcal{L}_L(w_L) = \frac{1}{N}\sum_{i=1}^{N} \| L^{(i)} - \overline{L} \|^2 = \overline{Var_L^2} \quad (8)$$

As a result, a simple equation is obtained for the loss function (10). Therefore, we consider the average log of the tasks' loss functions instead of the mean of the loss functions themselves.

$$\mathcal{L}(w) = \frac{1}{2}\Big(\log\big(\mathcal{L}_S(w_S)\big) + \log\big(\mathcal{L}_L(w_L)\big)\Big) \quad (9)$$

The new loss function aims to improve the validity of the body part predictions by capturing task-based uncertainty without changing the complexity of the model.

Experiments

In this section, we first introduce the datasets and evaluation metrics and then report the experimental results and analyze them.

A. Datasets and Metrics

We conduct the experiments on the COCO keypoints 2014 and COCO keypoints 2017 datasets [48]. These datasets are the largest collection of multi-instance person keypoint annotations which has been widely used in many studies. COCO datasets consist of many challenging situations for multi-person pose estimation problems. 17 keypoints including 12 human body parts and 5 facial keypoints are localized in the COCO keypoints dataset. COCO keypoints 2014 consists of 83k training data and 41k test data and COCO keypoints 2017 consists of 118k training and 41k test data. The COCO training set consists of over 100K person instances labeled with over 1 million keypoints. We report the results on both versions of COCO keypoints.

The performance of the proposed method is evaluated based on the object keypoint similarity (OKS) which is defined in COCO evaluation [49]. The role of OKS is the same as the IoU in object detection. OKS measures the degree of match between real and predicted poses. It ranges from 0 to 1 which refers to poor to perfect match. The mean average recall (AR) and the mean average precision (AP) over 10 OKS thresholds are used as the main competition metrics. Moreover, we assess the methods by AP and AR over thresholds 0.5 and 0.75, which are indicated by AP^{50} and AR^{50}, and AP^{75} and AR^{75}, respectively. Besides, results per each body part are presented to have a better analysis.

The results are reported on two models: 1) CMU-Pose: a model whose architecture is similar to our proposed model, but uses the CMU-Pose loss function [14] which is the average of the two tasks' loss functions and does not capture uncertainty, 2) CertainPose: the proposed method which captures task-based uncertainty as a new loss function.

B. Results and Discussion

Both models, CertaionPose and CMU-Pose, are trained on COCO keypoints 2014 training data. The MultiSGD optimizer with a learning rate of 2e-5 is used and the size of each batch is 10 images. CMU-Pose and CertainPose are trained for 100 and 18 epochs respectively. Practically, CertainPose can be trained faster than CMU-Pose.

Table 1: Comparison between CertainPose and CMU-Pose by mean of AP and AR metrics over all body parts on COCO validation sets 2014 and 2017

DB	Methods	AP	$\boldsymbol{AP^{50}}$	$\boldsymbol{AP^{75}}$	AR	$\boldsymbol{AR^{50}}$	$\boldsymbol{AR^{75}}$
Val2014	CMU-Pose	**0.59**	0.792	0.637	0.623	0.806	0.664
	CertainPose	0.589	**0.802**	**0.643**	**0.626**	**0.816**	**0.671**
Val2017	CMU-Pose	**0.578**	0.78	**0.625**	0.613	0.795	0.654
	CertainPose	0.575	**0.79**	0.624	**0.614**	**0.804**	**0.66**

The results of the test procedure for CMU-Pose and CertainPose on both datasets are shown in Table 1. The higher AP values refer to the more precise localization, and the higher ARs show more valid predictions. The results are measured by mean AP and mean AR over three values of OKS thresholds (0.5, 0.75, and 0.05:0.95) for all body parts. The higher threshold value considers the more perfect match between the estimated and real parts locations. The results show that CertainPose: 1) improves

AR measure definitely, 2) improves AP measure with lower OKS, and 3) has AP factor comparable to the base model when OKS is increased.

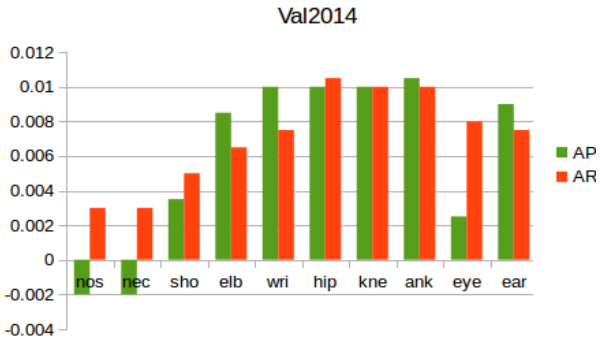


Fig. 3: The improvements of CertainPose for different keypoints on Val2014 dataset by $AP^{50}$and AR^{50}metrics for OKS=0.5.

Here, we explain the conclusions more clearly. First, CertainPose improves the AR measure by capturing task-based uncertainty through the loss function. This results in more valid and more certain outputs. In other words, the false positive rate is reduced in this method. Second, the AP measure improves because of training the two tasks fairer and predicting more accurate PAFs, which are impressive in keypoint association and body pose estimation. Third, while the CertainPose AP measure improves for lower OKS, it is not better than CMU-Pose for the higher OKS threshold, e.g. 0.95. It means that our model estimates more valid and more accurate body poses, but it fails to localize body parts more precisely since we apply a log operator over the distance between the predicted and the actual outputs in the loss function.

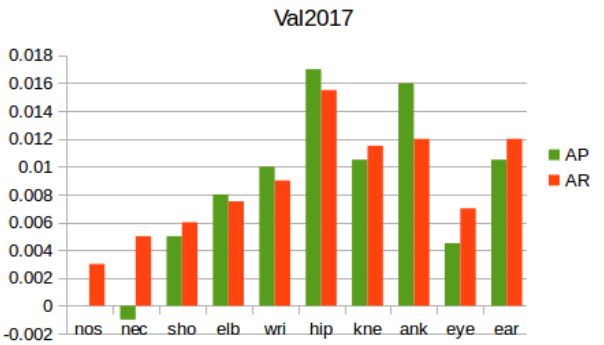


Fig. 4: The improvements of CertainPose for different keypoints on Val2017 dataset by $AP^{50}$and AR^{50}metrics for OKS=0.5.

CertainPose can associate the localized keypoints more accurately because PAFs are predicted more precisely. In the CMU-Pose loss function, equal weights are considered for PAFs and Confidence maps, but CertainPose considers task-dependent uncertainties to weigh the sub-loss functions. PAFs are more difficult than confidence maps to predict. Therefore, PAFs need higher weights and CertainPose predicts PAFs more precisely. Fig. 4 and Fig. 5 show the improvements of CertainPose for each keypoint in comparison with CMU-Pose on COCO val2014 and val2017 datasets. The keypoints like the elbows, hips, ankles, and knees which are connected with more clear limbs are localized more precisely.

We further show qualitative results for some images in Fig. 6. The (a) and (b) parts show the results of CertainPose and CMU-Pose, respectively. Some keypoints such as knees and elbows are predicted more accurately by the CertainPose method. Predicting the right elbow and left knee causes the more correct poses in the first two of the above images. Other shown samples demonstrate the power of CertainPose in PAF estimation. Higher accuracy in PAFs estimation is the reason of correct poses in the left hand and the left leg of the men, and the left hand of the baby in other three images, respectively.

We have analyzed the cases where our approach fails. Fig. 5. shows an overview of some failure cases and compares with the base method. The low resolution is the main cause of errors in the joint localization and PAFs estimation.

Realtime estimation is an important characteristic of HPE models in many real-world applications. CMU-Pose is the popular realtime multi-person pose estimation method. Its speed is independent of the number of people in the image. CertainPose improves the base model without adding any parameters. The main contribution of CertainPose is improving the CMU-Pose accuracy without decreasing the speed and increasing the complexity.

Table 2 shows the almost equal time of the two methods when perform single-scale process with the CUDA toolkit.

Table 2: Comparison between CertainPose and CMU-Pose by process time (ms)

Methods	CertainPose	CMU-Pose
CUDA (ms)	88.74	86.82

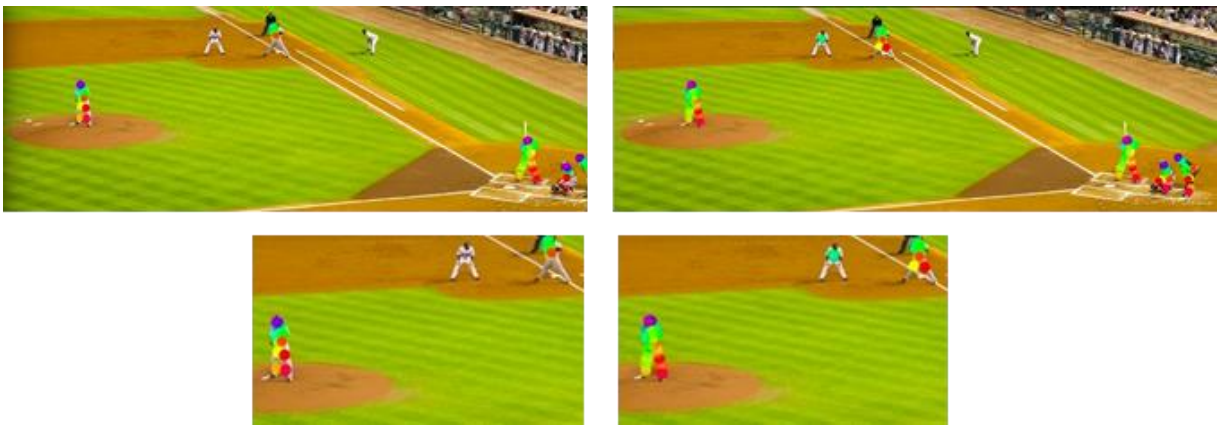
The goal of this research was to introduce the new loss function and investigate its performance in an applicable network, 2D HPE.

We show that we can increase the accuracy with the same number of parameters and inference speed. It is true that the improvement is not very significant, but one should note that the cost is not increased, either. In addition, the task-dependent uncertainties are captured and a few epochs are needed in the training step. The comparison of CertainPose and some other studies on COCO val split is shown in Table 3.

(a)

(b)

Fig. 6: The qualitative results of (a) CertainPose and (b) CMU-Pose.



(a)

(b)

(c)

Fig. 5: Visualization of some failure results of CertainPose on the COCO dataset and comparison between CertainPose (left) and CMU-Pose (right).

Osokin [16] introduces a lightweight OpenPose with fewer parameters and lower complexity compared to OpenPose. Cao et al. [14], the winner of the COCO 2016 keypoints challenge reports 58.4 AP for a model that is similar to CertainPose but, with two less layers. Newell et al. [50] propose a new approach for detections and group assignments. As reported in [51], their AP on COCO dataset is 56.9. Kocabas et al. [51] improve the base method by using a new grouping idea to associate body joints.

Table 3: Comparison of CertainPose and other works.

Methods	Osokin 161 [16]	Cao et al. [14]	Newell et al. [50]	Kocabas et al. [51]	CertainPose
AP	48.6	58.4	56.9	59.2	58.9

In summary, we compare the proposed idea and the baseline in Table 4. However, CMU-Pose is trained for 100 epochs, CertainPose needs 18 epochs. The runtime and number of parameters are almost the same.

The AP and AR comparisons show that CertainPose estimates more valid and accurate poses, and finds the less precise location for keypoints.

Conclusion

To obtain a more certain realtime multi-person pose estimation network, we propose a method to capture task-dependent uncertainties across the loss functions without increasing the number of parameters. As comparison Table 4 shows, the experiments prove that CertainPose: 1) needs fewer epochs for training, 2) preserves the realtime pose estimation property, 3) provides more valid and accurate estimations, and 4) locates keypoints less precisely.

In future work, we intend to examine different tasks and information to improve multi-person pose estimation.

The main weakness of our work is focusing on PAFs which causes less precise predicted heatmaps, particularly for keypoints with lower resolution (Fig. 5). We can use high resolution architecture instead of predictor units.

Also, the CertainPose idea can be the base method for incorporating other information to improve the pose estimation accuracy.

Table 4: Comparison between CertainPose and CMU-Pose.

Method	**Epochs**	**Runtime**	**Parameters**	AP^{50} **(V14)**	AP^{50} **(V17)**	AR^{50} **(V14)**	AR^{50} **(V17)**
CertainPose	18	88.74	44,970,966	0.802	0.79	0.816	0.804
CMU-Pose	100	86.82	44,970,966	0.792	0.78	0.806	0.795

Author Contributions

Z. Ghasemi-Naraghi implemented the proposed model and performed the experiments. Z. Ghasemi-Naraghi and A. Nickabadi interpreted the results. Z. Ghasemi-Naraghi, A. Nickabadi and R. Safabakhsh wrote the manuscript.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

HPE Human Pose Estimation
HOG Histogram of Oriented Gradients
CNN Convolutional Neural Network
CPM Convolutional Pose Machine
PAFs Part Affinity Fields
MTL Multi-Task Learning
PSM Pictorial Structure Model
RPN Region Proposal Network
RIE Resolution Irrelevant Encoding
DBL Difficulty Balanced Loss
PIF Part Intensity Field
PAF Part Association Field
CNN Convolutional Neural Networks
MTL Multi-Task Learning
OKS Object Keypoint Similarity
AR Average Recall
AP Average Precision

References

[1] Y. Yang, D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in Proc. CVPR 2011: 1385–1392, 2011.

[2] P. F. Felzenszwalb, D. P. Huttenlocher, "Pictorial structures for object recognition," Int. J. Comput. Vision, 61(1): 55–79, 2005.

[3] X. Chen, A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in Proc. NIPS, 2014.

[4] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 770–778, 2016.

[5] S. Ren, K. He, R. Girshick, J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Proc. NIPS, 2015.

[6] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, “Towards accurate multi-person pose estimation in the wild,” in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 4903–4911, 2017.

[7] K. He, G. Gkioxari, P. Dollár, R. Girshick, “Mask r-cnn,” in Proc. the IEEE International Conf. on Computer Vision: 2961–2969, 2017.

[8] S. E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, ” Convolutional pose machines,” in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 4724–4732, 2016.

[9] H. S. Fang, S. Xie, Y. W. Tai, C. Lu, “Rmpe: Regional multi-person pose estimation,” in Proc. the IEEE International Conference on Computer Vision: 2334–2343, 2017.

[10] L. Ladicky, P. H. Torr, A. Zisserman, “Human pose estimation using a joint pixel-wise and part-wise formulation,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition: 3578– 585, 2013.

[11] U. Iqbal, J. Gall, “Multi-person pose estimation with local joint-to-person associations,” in Proc. European Conference on Computer Vision: 627–642, 2016.

[12] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, B. Schiele, “Deepcut: Joint subset partition and labeling formulti person pose estimation,” in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 4929–4937, 2016.

[13] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, “Deepercut: A deeper, stronger, and faster multi-person pose estimation model,” in Proc. European Conference on Computer Vision: 34–50, 2016.

[14] Z. Cao, T. Simon, S. E. Wei, Y. Sheikh, “Realtime multi-person 2dpose estimation using part affinity fields,” in Proc. the IEEE Conf. Computer Vision and Pattern Recognition: 7291–7299, 2017.

[15] X. Zhu, Y. Jiang, Z. Luo, “Multi-person pose estimation for posetrack with enhanced part affinity fields,” in Proc. ICCV PoseTrack Workshop, 7, 2017.

[16] D. Osokin, “Real-time 2d multi-person pose estimation on cpu: Lightweight openpose,” in Proc. the 8th International Conference on Pattern Recognition Applications and Methods: 744–748, 2019.

[17] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, Y. Sheikh, “Openpose: real-time multi-person 2d pose estimation using part affinity fields,” IEEE Trans. Pattern Anal. Mach. Intell., 43(1): 172–186, 2019.

[18] OpenPose library. https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[19] H. Liu, D. Luo, S. Du, T. Ikenaga, “Resolution irrelevant encoding and difficulty balanced loss based network independent supervision for multi-person pose estimation,” in Proc. 13th International Conf.HumanSystem Interaction (HSI):112–117, 2020.

[20] G. H. Martınez, “OpenPose: Whole-Body Pose Estimation,” April, 2019.

[21] T. Gong, T. Lee, C. Stephenson, V. Renduchintala, S. Padhy, A. Ndirango, G. Keskin, O. H. Elibol, “A comparison of loss weighting strategies for multi task learning in deep neural networks,” IEEE Access, 7: 141627–141632, 2019.

[22] Q. Dang, J. Yin, B. Wang, W. Zheng, “Deep learning based 2d human pose estimation: A survey,” Tsinghua Sci. Technol., 24(6): 663–676, 2019.

[23] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, M. Shah, “Deep learning-based human pose estimation: A survey,” arXiv preprint arXiv:2012.13392, 2020.

[24] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, C. Yang, “The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation,” IEEE Access, 8: 133330–133348, 2020.

[25] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, E. H. Zahzah, “Human pose estimation from monocular images: A comprehensive survey,” Sensors, 16(12): 1966, 2016.

[26] G. Rogez, C. Schmid, “Mocap-guided data augmentation for 3d pose estimation in the wild,” in Proc. Advances in Neural Information Processing Systems (NIPS): 3108–3116, 2016.

[27] H. Jiang, “Finding human poses in videos using concurrent matching and segmentation,” in Proc. Asian Conference on Computer Vision: 228–243, 2010.

[28] H. Sidenbladh, F. De la Torre, M. J. Black, “A framework for modeling the appearance of 3d articulated figures,” in Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580): 368–375, 2000.

[29] J. J. Tompson, A. Jain, Y. LeCun, C. Bregler, “Joint training of a convolutional network and a graphical model for human pose estimation,” in Proc. Advances in Neural Information Processing Systems, 27, 2014.

[30] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, Y. Sheikh, “Pose machines: Articulated pose estimation via inference machines,” in Proc. European Conf. Computer Vision: 33–47, 2014.

[31] MSCOCO Dataset, https://cocodataset.org/#home.

[32] T. Simon, H. Joo, I. Matthews, Y. Sheikh, “Hand keypoint detection in single images using multiview bootstrapping,” in Proc. the IEEE Conf. Computer Vision and Pattern Recognition: 1145–1153, 2017.

[33] S. Kreiss, L. Bertoni, A. Alahi, “Pifpaf: Composite fields for human pose estimation,” in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 11977–11986, 2019.

[34] N. Nakano, T. Sakura, K. Ueda, L. Omura, A. Kimura, Y. Iino, S. Fukashiro, S. Yoshioka, “Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras,” Fron. sports Active Living, 2(50), 2020.

[35] N. D. Reddy, L. Guigues, L. Pishchulin, J. Eledath, S. G. Narasimhan, “Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking,” in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 15190–15200, 2021.

[36] N. D. Reddy, M. Vo, S. G. Narasimhan, “Occlusion-net: 2d/3d occluded keypoint localization using graph networks,” in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 7326–7335, 2019.

[37] Y. Cheng, B. Wang, B. Yang, R. T. Tan, “Monocular 3d multi-person pose estimation by integrating top-down and bottom-up networks,” in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 7649–7659, 2021.

[38] H. Tu, C. Wang, W. Zeng, “Voxelpose: Towards multi-camera 3d human pose estimation in wild environment,” in Proc. 16th European Conference on Computer Vision–ECCV 2020, Part I 16: 197–212, 2020.

[39] G. Zhang, J. Liu, H. Li, Y. Q. Chen, L. S. Davis, “Joint human detection and head pose estimation via multistream networks for rgb-d videos,” IEEE Signal Process. Lett., 24(11): 1666–1670, 2017.

[40] M. Schwarz, H. Schulz, S. Behnke, “Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features,” in Proc. 2015 IEEE International Conference on Robotics and Automation (ICRA): 1329–1335, 2015.

[41] A. Krull, E. Brachmann, F. Michel, M. Y. Yang, S. Gumhold, C. Rother, “Learning analysis-by-synthesis for 6d pose estimation in rgb-d images,” in Proc. the IEEE International Conference on Computer Vision: 954–962, 2015.

[42] Y. Gal, “Uncertainty in deep learning,” University of Cambridge 1(3), 2016.

[43] A. Kendall, Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?” in Proc. Advances in Neural Information Processing Systems: 5574–5584, 2017.

[44] F. K. Gustafsson, M. Danelljan, T. B. Schon, "Evaluating scalable bayesian deep learning methods for robust computer vision," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops: 318–319, 2020.

[45] I. Misra, A. Shrivastava, A. Gupta, M. Hebert, "Cross-stitch networks for multi-task learning," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 3994–4003, 2016.

[46] Z. Chen, V. Badrinarayanan, C. Y. Lee, A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in Proc. International Conference on Machine Learning: 794–803, 2018.

[47] A. Kendall, Y. Gal, R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 7482–7491, 2018.

[48] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, "Microsoft coco: Common objects in context. In Proc. European Conf. Computer Vision: 740–755, 2014.

[49] M. Ruggero Ronchi, P. Perona, "Benchmarking and error diagnosis in multi-instance pose estimation," in Proc. the IEEE International Conference on Computer Vision: 369–378, 2017.

[50] A. Newell, Z. Huang, J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in Proc. Advances in Neural Information Processing Systems: 2278–2288, 2017.

[51] M. Kocabas, S. Karagoz, E. Akbas, "Multiposenet: Fast multi-person pose estimation using pose residual network," in Proc. the European Conference on Computer Vision (ECCV): 417–433, 2018.

Biographies

Zeinab Ghasemi-Naraghi is a Ph.D. student in Artificial Intelligence at the Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran. She earned a M.S. degree in Artificial Intelligence from Sharif University of Technology and received the B.S. degree in Computer Engineering from Shahid Beheshti University in 2013 and 2009, respectively. Her research interests are mainly in computer vision, deep learning and probabilistic graphical models.

- Email: z_naraghi@aut.ac.ir
- ORCID: 0009-0008-9545-8706
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Ahmad Nickabadi received the B.S. degree in Computer Engineering and the M.S. and Ph.D. degrees in Artificial Intelligence from the Amirkabir University of Technology (AUT), Tehran, Iran, in 2004, 2006, and 2011, respectively. Since 2012, he has been an Assistant Professor with the Computer Engineering Department, AUT. His research interests include the analysis of image and video content using deep learning and probabilistic graphical models with a special focus on activity recognition, face recognition, and face synthesis.

- Email: nickabadi@aut.ac.ir
- ORCID: 0000-0003-3709-1041
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://aut.ac.ir/cv/2387/Ahmad%20Nikabadi

Reza Safabakhsh received the B.S. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 1976 and the M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Tennessee, Knoxville, in 1980 and 1986, respectively. He worked at the Center of Excellence in Information Systems, Nashville, TN, USA, from 1986 to 1988. Since 1988, he has been with the Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran, where he is currently a professor and the director of the Computer Vision Laboratory. His current research interests include neural networks, computer vision, and deep learning. Dr. Safabakhsh is a member of the IEEE and several honor societies, including Phi Kappa Phi and Eta Kapa Nu. He was the founder and a member of the Board of Executives of the Computer Society of Iran, and was the President of this society for the first 4 years.

- Email: safa@aut.ac.ir
- ORCID: 0000-0002-4937-8026
- Web of Science Researcher ID:
- Scopus Author ID
- Homepage: https://aut.ac.ir/cv/2455/REZA%20SAFABAKHSH

How to cite this paper:
Z. Ghasemi-Naraghi, A. Nickabadi, R. Safabakhsh, "Multi-Task learning using uncertainty for realtime multi-person pose estimation," J. Electr. Comput. Eng. Innovations, 12(1): 147-162, 2024.

DOI: 10.22061/jecei.2023.9848.657

URL: https://jecei.sru.ac.ir/article_1985.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Text Detection and Recognition for Robot Localization

Z. Raisi [1,2,*], *J. Zelek* [2]

[1]*University of Waterloo, Waterloo, Canada and Chabahar Maritime University, Chabahar, Iran.*

[2]*Systems Design Engineering Department, University of Waterloo, Canada.*

Article Info

Article History:
Received 26 June 2023
Reviewed 13 August 2023
Revised 05 September 2023
Accepted 08 September 2023

Keywords:
Text detection
Text recognition
Robot localization
Deep learning
Visual place recognition

*Corresponding Author's Email Address: *zraisi@uwaterloo.ca*

Abstract

Background and Objectives: Signage is everywhere, and a robot should be able to take advantage of signs to help it localize (including Visual Place Recognition (VPR)) and map. Robust text detection & recognition in the wild is challenging due to pose, irregular text instances, illumination variations, viewpoint changes, and occlusion factors.

Method: This paper proposes an end-to-end scene text spotting model that simultaneously outputs the text string and bounding boxes. The proposed model leverages a pre-trained Vision Transformer (ViT) architecture combined with a multi-task transformer-based text detector more suitable for the VPR task. Our central contribution is introducing an end-to-end scene text spotting framework to adequately capture the irregular and occluded text regions in different challenging places. We first equip the ViT backbone using a masked autoencoder (MAE) to capture partially occluded characters to address the occlusion problem. Then, we use a multi-task prediction head for the proposed model to handle arbitrary shapes of text instances with polygon bounding boxes.

Results: The evaluation of the proposed architecture's performance for VPR involved conducting several experiments on the challenging Self-Collected Text Place (SCTP) benchmark dataset. The well-known evaluation metric, Precision-Recall, was employed to measure the performance of the proposed pipeline. The final model achieved the following performances, Recall = 0.93 and Precision = 0.8, upon testing on this benchmark.

Conclusion: The initial experimental results show that the proposed model outperforms the state-of-the-art (SOTA) methods in comparison to the SCTP dataset, which confirms the robustness of the proposed end-to-end scene text detection and recognition model.

Introduction

We live in a visual world; signage is everywhere. Whether it is a street sign, a billboard, a house or room number, or labels such as a license plate or a person's name, signage provides us useful information in terms of location and identity. There have been many classifiers developed that are able to identify street signs or license plates with highly constrained priors on the method that do not allow their extension to general text in the wild detection and recognition.

However, to take advantage of all the signage available, we need to be able to detect signage (i.e., text) anywhere (i.e., in the wild). OCR is a well-solved problem for text detection and recognition in highly constrained environments; however, detecting and recognizing text anywhere is a challenging problem.

Signage can help a robot localize or map an environment. Typically, for SLAM processes, direct (i.e., pixel) or indirect features are used. Signage can provide a coarse localization globally when the signage indicates an

address or location. Also, the letters and numbers in a sign and their perspective can be used to determine relative pose if it can be assumed the signage is on a planar surface or even just vertical with respect to the ground plane. Visual Place Recognition (VPR) [15], [22], [41], [44], [69] aims to aid a vision-guided system to localize with respect to a previously visited place. VPR has uses in loop closure detection for visual SLAM and localization in general. Challenges in VPR include appearance variation due to perceptual aliasing, illumination, viewpoint changes, pose, weather, and seasons, to name a few. Most techniques are focused on features (i.e., indirect) [29] and sets of these features (e.g., BOW Bag of Words) methods.

Text spotting in wild images is also called end-to-end scene text detection and recognition [40], [52]. Simultaneous text detection and recognition go hand in hand. In scene text detection, the goal is to localize words in the image, and for scene text recognition, the aim is to convert the patch of cropped word images into a sequence of characters. Like scene text detection and recognition tasks, scene text spotting also encounters different challenging problems, including irregular text, illumination variations, low-resolution text, occlusion, *etc* [50].

Previous methods in scene text detection and recognition have utilized a convolutional neural network (CNN) as a feature extractor [19], [37], [56], [57] and Recurrent Neural Networks (RNN) [4], [21], [59] for capturing sequential dependency. Despite achieving promising performances on various challenging benchmark datasets [9], [16], [23]-[26], [35]-[42], [46], [48], [58], [60], [65]-[67], it has been shown that there are two main challenges for detecting or recognizing text in the wild images that have been studied in the past years. (1) Irregular text refers to text with arbitrary shapes that usually have severe orientation and curvature, and (2) occlusion, which makes poor performance on the existing methods [4], [5], [61] due to their reliance on the visibility of the target characters in the given images. Furthermore, CNNs have two significant drawbacks: (1) they have problems in capturing long-range dependencies (e.g., arbitrary relations between pixels in spatial domains) due to their fixed-size window operation [70], (2) they suffer from dynamical adoption to the changes to the inputs because the convolution filter weights are tuned to a specific training distribution [27].

Recent end-to-end scene text spotting methods [28], [54], [55], [58] utilized transformers [64] in their architecture and achieved superior performance in many benchmarks [9], [67]. Transformers [64] and their variations [7], [10], [70] are a new deep-learning architecture that mitigates the issues mentioned above for CNNs. Unlike Recurrent Neural Networks (RNNs), transformers are models that learn how to encode and decode data by looking not only backward but also forward to extract relevant information from a whole sequence, allowing conducting complex tasks such as machine translation [64], speech recognition [8], and recently, computer vision [7], [12], [27]. The attention mechanism allows the transformers to reason more effectively and focus on the relevant parts of the input data (e.g., a word in a sentence for machine translation and a character of a word in a text image for detection and recognition) as needed.

Visual place recognition (VPR) [22] aims to recognize previously visited places using visual information with resilience to perceptual aliasing, illumination, and viewpoint changes. Most of the techniques in VPR used keypoint features such as corners, edges, or blobs to represent and match distinctive points between the images. However, many keypoint features are needed to extract from images to establish a robust and repeatable representation of places and facilitate reliable localization and mapping in various applications like autonomous navigation, augmented reality, and robot localization. This process can be expensive in terms of computation and matching. On the other hand, as shown in Fig. 1, text features are semantic indexes and fewer in number compared to point features. Text instances that appear in the wild images, such as street signs, billboards, and shop signage, usually carry extensive discriminative information. VPR task can take advantage of these scene texts with high-level information for previously visited place recognition.

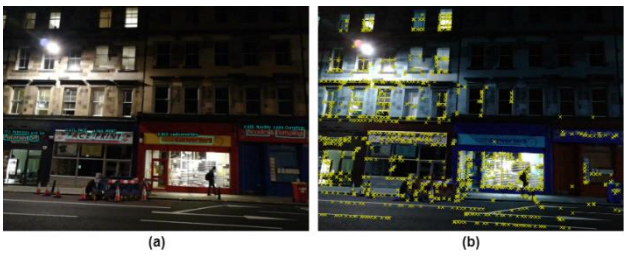


Fig. 1: Comparing the (a) Text features used in the proposed E2E text detection model and (b) Key point features used in different VPR techniques [11], [45]. Text features are shown with 'cyan' color boxes, and the 'x' marks with 'yellow' color denote the keypoint features.

This paper leverages a pre-trained end-to-end transformer-based text spotting framework for the VPR task. Unlike [22], which used two separate modules of detection and recognition for extracting the text regions, the final model can directly read the text instances from the given frame in an end-to-end manner. Furthermore, by equipping a masked autoencoder (MAE) [18] as a backbone, the proposed model is more robust in

capturing occluded text instance regions, which makes it more suitable for visual place recognition and other applications, including assistive technology for visually impaired people, autonomous vehicles, automated translation, and language processing in the wild images, information extraction from videos, and mobile applications for OCR and text, to name a few [36], [50].

The main contributions of this paper are as follows: (1) it utilizes an end-to-end transformer-based scene text spotting pipeline for the VPR application for the first time. The main difference between this method and other approaches is that it can directly output semantic text features (word instances and their bounding boxes), much less than the keypoint features used in most VPR techniques. (2) The proposed model utilizes a modified version of ViT by leveraging masked as input and adding a multi-scale adapter at the output to extract suitable features later for detection and recognition. (3) At the final stage, after utilizing a transformer-based detection architecture, this work uses a prediction head capable of simultaneously detecting the characters in the input image with their predicted classes and the bounding boxes of the word instances in the image. (4) By joining the middle point of the detected characters and their classes, the proposed model can handle arbitrary shapes text and output polygon bounding boxes with their word instances simultaneously, which makes it suitable not only for VPR but for other several text detection and recognition based in the wild applications. (5) This work also provides several quantitative and qualitative comparisons of the proposed technique with state-of-the-art (SOTA) in both VPR and scent text techniques.

Related Work

Scene text spotting aims to detect and recognize text instances from a given image end-to-end [14], [31], [33], [38], [39], [43], [47], [57]. Like different computer vision tasks, deep learning techniques using CNN/RNN-based methods and transformer-based methods are dominant frameworks in scene text spotting.

Early methods [31], [38] in scene text spotting have mainly utilized a deep-learning convolutional neural network (CNN) as a feature extractor [20] and Recurrent Neural Networks (RNN) [4], [21], [59] to read horizontal scene text. For example, Li *et al.* [31] combined the detection and recognition framework to present the first text-spotting method by using a shared CNN backbone encoder, followed by RoIPooling [57] as detection. Then, the resulting features are fed into the RNN recognition module to output the final word instances for a given input image. FOTS [38] utilized an anchor-free CNN-based object detection framework that improved both the training and inference time. It also uses RoIRotate for reading rotated text instances.

Since the text in the wild images appears in arbitrary shapes, including multi-oriented and curved, several methods [14], [33], [39], [47] targeted reading these types of text instances. These methods usually used a CNN-based segmentation network with multiple post-processing stages to output polygon box coordinates for the final irregular texts. For instance, in [47], a RoiMask is used to connect both the detection and recognition modules for capturing arbitrarily shaped text. Liu [39] leveraged a Bezier curve representation for the detection part, followed by a Bezier Align module to rectify the curved text instances into a regular text before feeding it to the attention-based recognition part. Some methods [6], [55] targeted spotting individual characters and merging them to output the final arbitrary shape text instance.

The Transformer framework, introduced by Vaswani et al. [64] for natural language processing tasks, has become a foundational architecture in various domains, including computer vision. In natural language processing, the Transformer architecture was initially designed to handle sequential data, such as sentences. It introduces a novel self-attention mechanism that allows the model to weigh the importance of different input elements when generating outputs. This attention mechanism enables parallel processing of sequences and mitigates the limitations of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) in handling long-range dependencies.

Inspired by the success of Transformers in natural language processing, Dosovitskiy et al. [12] extended the Transformer framework to computer vision tasks with the Vision Transformer (ViT) architecture. ViT treats images as sequences of non-overlapping patches, which are then flattened into 1D sequences to be processed by the Transformer. By leveraging self-attention, ViT captures global contextual information from the entire image and allows for efficient modeling of long-range dependencies.

The self-attention mechanism is the cornerstone of the Transformer framework. It computes a weighted sum of values (representations of input elements) based on their relevance to a query (a representation to be updated). Self-attention computes attention scores between each pair of elements in the input sequence and generates attention weights that signify the importance of different elements relative to each other. This attention mechanism allows the model to adaptively focus on relevant input parts during each processing step, facilitating better representation learning [64].

In the Transformer architecture, each self-attention layer is followed by a feed-forward neural network module. The feed-forward module consists of two linear transformations separated by a non-linear activation function, typically a GELU (Gaussian Error Linear Unit) or ReLU (Rectified Linear Unit). This module introduces non-

linearity and enables the model to learn complex relationships between different elements in the input sequence. Combining self-attention and feed-forward modules empowers the Transformer to effectively model local and global dependencies within the input sequence, making it highly adaptable to various tasks, including computer vision [12].

Recently, with the advancement of transformers [64] in computer vision fields [17], [27], [63], several SOTA scene text spotting methods [3], [13], [30], [49], [51], [53] proposed to take the benefit of transformer-based pipelines in their framework. These methods achieved superior performance in both regular and irregular benchmark datasets. For example, Kittenplon et al. [28] utilized a transformer-based detector, Deformable-DETR [70], as its primary framework by proposing a multi-task prediction head that can output word instances and box coordinates of an arbitrary shape text. [68] used transformers as the main block for an end-to-end text-spotting framework for text detection and recognition in wild images. These methods removed the dependency of region-of-interest operations and post-processing stages in their framework. Thus, they can output both Bezier curve and polygon representations and achieve superior benchmark performance. Very recently, Raisi et al. [54] proposed an end-to-end framework for scene text spotting that is also capable of improving the recognition performance for an adverse situation like occlusion. This method utilized an MAE in their pipeline equipped with a powerful detector, namely Deformable-DETR [70], to capture the arbitrary shape of occluded text instances in the wild images. In this study, a pre-trained model from [54] is utilized for the VPR task.

The Proposed Scene Text Spotting Architecture

For complete text reading, simultaneous text detection and recognition are required. Unlike stepwise detection and recognition, as utilized in [22], the end-to-end framework will improve the overall speed by eliminating multiple processing steps. Furthermore, an end-to-end transformer is expected to offer higher accuracy compared to previous end-to-end CNN-based approaches [38], [39].

A. Backbone

The overall framework of the proposed method is shown in Fig. 2 Inspired [32], the proposed model uses pre-trained models of the Vision Transformer architecture (ViT) [12] as the backbone. The 2-Dimensional (2D) input image ($I \epsilon \mathbb{R}^{H\times W\times C}$) is first split into a non-overlapping sequence of patches ($I' \epsilon \mathbb{R}^{N\times P^2\times C}$), where ($H$,$W$) represent the height and width of the image, C is the number of channels and (P, P) denote the resolution of the patches. The number of patches ($N = HW/P^2$) is set to 16.

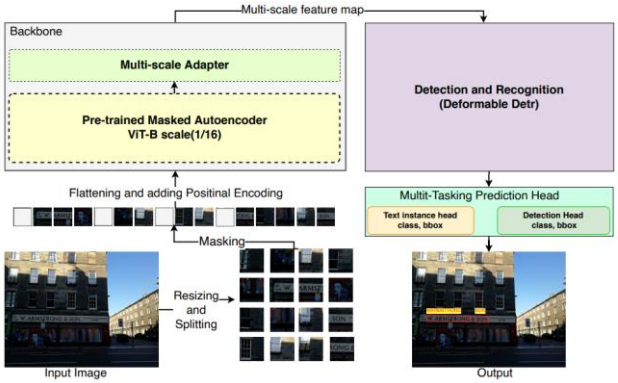


Fig. 2: Block diagram of the proposed scene text spotting architecture using a transformer for VPR [54]. Unlike the stepwise pipeline in [22], the proposed model outputs the bounding box coordinates and the word instances in an end-to-end manner for the VPR task. It is best viewed when zoomed in.

After masking a large set of the input patches ($\sim 75\%$) and adding the 1D position embedding, these patches are passed into the encoder of the MAE ViT block, which contains several multi-head self-attention and feed-forward modules. The encoder operates on unmasked patches to acquire the visual feature embeddings. However, the final output of the ViT encoder backbone is single-scale due to the columnar structure of ViT, which makes them inadequate for detecting multi-scale text instances. To address this, a multi-scale adapter module [32] is utilized. It is worth mentioning that the model uses a pre-trained MAE [18] (ViT-Base/16) as the backbone for feature extraction. This backbone was further fine-tuned on 36 classes of alphanumeric characters (More details in subsection A of Experimental Results).

B. Multi-scale adapter

Inspired by [32], [54], a single-scale ViT into the multi-scale FPN for capturing different resolutions of text regions is adapted. The multi-scale feature map module utilizes the idea of up-sampling or down-sampling into the intermediate single-scale ViT's feature map with a columnar structure [32]. The Multi-scale Adapter module in Fig. 2 consists of 4 up-sampling and down-sampling subblocks. The output feature map of the first block undergoes up-sampling with a scaling factor of 4. Subsequently, the output of the following block is up-sampled, but this time by a factor of 2. On the other hand, the output of the third block remains unchanged, which remains equal to the original feature map. Finally, the output feature map of the last block undergoes down-sampling with a scaling factor of 2. As a result of these operations, a set of multi-scale features is obtained, containing feature maps with different resolutions.

These multi-scale feature maps are then fed into an extended detector [70], which utilizes this information to perform text detection and recognition. By leveraging the diverse scales and resolutions captured in the multi-scale

features, the upgraded detector can effectively handle text instances of varying sizes and efficiently analyze the input image at different levels of detail.

C. Text Predictor

After feature extraction and multi-scaling, the resulting feature maps are fed to the text of the final module to detect and recognize the text instance of a given image. As shown in Fig. 2, the proposed text predictor leverages a modified Deformable-DETR [70] with a multi-task prediction head. This work introduces an enhanced adaptation of the FFN layer, which differs from the [70] architecture. The proposed modifications aim to significantly enhance its ability to capture the distinctive text features produced by the encoder's Multi-Head Self Attention (MHSA) mechanism. The improved FFN now comprises two layers of 1×1 convolutions, supplemented by ReLU activations, and ultimately integrated with a residual connection. This innovative approach effectively amplifies the FFN's capacity to encapsulate and process essential information from the encoder's MHSA, resulting in a more robust and efficient Transformer model.

During training, the encoder's multi-head self-attention detector learns how to separate individual characters and word instances in the scene image by performing global computations. The decoder typically learns how to attend to a different part of characters in words by using different learnable vectors (so-called object queries). During training, the multi-task head (last layer of the decoder) can directly predict both absolute bounding box coordinates and sequence of characters, eliminating the use of any hand-designed components and post-processing like anchor design and non-max suppression. To achieve this, a novel loss function based on optimal bipartite matching between the predicted text instances and the corresponding ground truth is leveraged. This matching process is crucial as it allows us to establish one-to-one correspondences and is efficiently computed using the Hungarian algorithm [70] explicitly adapted for this task. Using the Hungarian algorithm, the model can determine the optimal matching between the predicted and ground-truth elements. This matching information is instrumental in evaluating the performance of the Transformer model for character and word prediction within the text regions. The final loss function takes advantage of these optimal correspondences, enabling the model to learn and improve its predictions more effectively, enhancing accuracy and performance in the task. This work implements the same text filtering criteria introduced in [22] for comparing the query and inference frames.

Experimental Results

A. Implementation Details

The final model is trained on 4 GPUs of NVidia A100. First, by using about 500K cropped alphanumeric synthetic character images from the SynthText dataset [16] for 20 epochs are used to train the pre-trained encoder backbone of MAE (ViT-Base/16) [22] to make it more appropriate for scene text detection and recognition application for 200 epochs. Subsequently, 300 images of ICDAR15 [25] datasets are combined to fine-tune the final model. The Deformable DETR [70] module's object queries are set to 300, and the AdamW optimizer is used to optimize the model's parameters. The following augmentation strategies are applied to the input images during the learning process: horizontal and vertical flip, image resizing, brightness, contrast, and saturation. The model is trained with a batch size of 2 per GPU by employing a learning rate of 1×10^{-4} throughout the training process, and the whole process of training time takes ~ 23 hours. An NVIDIA RTX 3080Ti GPU with 12GB of memory is used for testing the final model.

B. Datasets

The **Self-Collected TextPlace (SCTP)** Dataset [22] is designed explicitly for visual place recognition tasks in urban places. The images of this dataset are captured using a side-looking mobile phone camera. These images include three pairs of map and query sequences in outdoor streets and an indoor shopping mall and contain significant challenging scenarios, including high dynamics, random occlusions, illumination changes, irregular text instances, and viewpoint changes.

The **ICDAR15** [25] is a challenging dataset that contains various indoor and outdoor multi-oriented text instances. Like most of the images in the VPR applications, this dataset has a wide variety of blurry and low-resolution text. The text instances in this dataset are annotated in quadrilateral bounding box annotations.

C. Evaluation Metrics

In the context of text detection and visual place recognition, Precision, Recall, and H-mean are widely used evaluation metrics to assess the performance of the trained model. Precision, recall, and H-mean are measurements that are accepted in almost all text detection communities [23]-[26], [35]-[42], [65]-[67]. These metrics are based on comparing the predicted text regions and the ground truth (manually annotated) text regions in an image, which can be described as follows:

Precision: measures the proportion of predicted text regions correctly identified as text regions among all the predicted text regions. It quantifies the model's ability to avoid false positives. The formula for precision is:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

where True Positives (TP) are the number of correctly predicted text regions, and False Positives (FP) are the number of non-text regions incorrectly predicted as text

regions.

Recall: measures the proportion of correctly predicted text regions out of all the ground truth text regions in the image. It assesses the model's ability to avoid false negatives. It can be defined as follows:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

where False Negatives (FN) are the number of text regions that the model did not correctly predict.

H-mean (Harmonic Mean): The H-mean, the F1-score, is the harmonic mean of precision and recall. It provides a balanced measure that considers both precision and recall simultaneously. The formula for H-Mean is:

$$H - Mean = 2\frac{Precison \times Recall}{Precison + Recall} \quad (3)$$

These metrics are essential in text detection evaluation [25], as they provide insights into the model's accuracy in correctly identifying text regions and its ability to balance false positives and false negatives. Researchers use these metrics to compare text detection algorithms and fine-tune models for optimal performance.

In this work, to compare the performance of the proposed model with SOTA VPR methods [1], [2], [22] [22], the same evaluation metrics, namely, precision-recall evaluation measurements are followed as in [22], [62].

Fig. 3: Query image and matching reference examples of [22] dataset. The proposed model detects and recognizes the most challenging text instances required to match the frames of the query (top column) and reference (bottom column). It is best viewed in color when zoomed in. The output results are indicated in 'cyan' color.

For end-to-end text detection and recognition that aims to output the correct string of the word instances in the image, in addition to the detected text metrics, the H-mean (F1-score) is used for the evaluation [3], [13], [14], [30], [31], [33], [38], [39], [43], [47], [49], [51], [57].

D. Quantitative Comparison of the VPR SCTP dataset with SOTA methods

The quantitative results of the proposed model with several SOTA methods [1], [2], [11], [22], [45] on the SCTP dataset [22] are shown in Table 1. The proposed model achieved the best performance in terms of recall for this dataset, which contains significant challenges like irregular and partially occluded text instances. This performance confirms the effectiveness of the proposed method for VPR.

Table 1: Precision-Recall comparison of the proposed model with SOTA methods including TextPlace [22], ToDayGAN [1], NetVLAD [2], SeqSLAM [45], and FAB-MAP [11] using SCTP [22] dataset. The best performance is highlighted in bold.

Model	Recall				
	0.2	0.4	0.6	0.8	0.9
Proposed model	1	1	1	**0.97**	**0.93**
TextPlace	1	1	1	0.96	0.91
NetVLAD-10	1	1	1	0.95	**0.93**
NetVLAD-20	1	1	1	0.91	0.87
NetVLAD-30	1	1	0.97	0.85	0.83
ToDayGAN-10	0.50	0.55	0.58	0.57	0.56
ToDayGAN-20	0.40	0.40	0.40	0.38	0.38
ToDayGAN-30	0.26	0.24	0.24	0.25	0.24
FAB-MAP-10	0.79	0.69	0.67	0.65	0.63
FAB-MAP-20	0.76	0.69	0.67	0.63	0.60
FAB-MAP-30	0.68	0.67	0.67	0.62	0.58
SeqSLAM	0.30	0.24	0.18	0.13	0.13

E. Qualitative Comparison on the VPR SCTP dataset

Fig. 3 illustrates the qualitative results of the proposed model on the SCTP [22] dataset. As seen, the model successfully read challenging text instances of both query and reference frames. The results of the proposed model are also compared with some of the SOTA techniques [2], [22], [45], as shown in Fig. 4; the proposed text spotting model correctly matches the query frame with frame in inference.

Fig. 4: Qualitative comparison of the proposed model with SOTA methods [2], [22], [45] on the SCTP dataset. The correct and incorrect results are bounded with green and red colors.

F. Quantitative Comparison with SOTA text detection and recognition approaches using ICDAR Dataset.

The proposed model also is compared with some SOTA scene text detection and recognition approaches [5], [34], [38], [71], [72], [74], [75]. As seen from Table 2, while these methods are trained on many images of synthetic

datasets and fine-tuned on real-world datasets, the proposed model achieves the best performance (P = 90.2) regarding precision for text detection and competitive performances in terms of recall and H-mean. It also performed well while testing for end-to-end text detection and recognition (E2E H-mean = 68.2). Since using the number of training images affects the final performance of deep learning models, for a fair comparison, only SOTA methods that used close images to the proposed model are selected for comparison.

Table 2: Quantitative comparison of the model with the baseline Textboxes++ [34] and other SOTA text detection and recognition methods using the ICDAR15 [25] dataset. P, R, H, and F mean Precision, Recall, H-mean, and F-measure, respectively. E2E denotes end-to-end text spotting, and FPS is Frames per second. The best and second-best performances are highlighted in bold and underlined.

Model	Detection			E2E	FPS
	P	R	H	F	
CRAFT [5]	88.5	**84.69**	**86.9**	--	--
PSENet [74]	86.9	84.5	85.6	--	--
EAST [75]	83.3	78.3	80.7	--	--
FOTS [38]	88.8	82.0	85.3	--	--
DRGN [71]	88.5	84.6	86.5	--	--
CharNetR50 [72]	--	--	--	<u>60.72</u>	--
Textboxes++[34]	87.8	78.5	82.9	51.9	2.3
Proposed Model	**90.2**	<u>83.1</u>	<u>86.5</u>	**68.2**	**11.0**

G. *Ablation Experiments*

1) Output Feature Comparison of the Proposed Model and VPR Methods. As mentioned in the Introduction, VPR algorithms mainly design their architecture to extract Point features, also known as keypoint features, to represent and match distinctive points between images for place recognition tasks. These algorithms detect and describe keypoint features based on local image information around the keypoints. To compare the output semantic features extracted from the proposed model and the keypoint features of VPR techniques, a qualitative ablation study is conducted, which is shown in Fig. 5; the text features of the proposed model are more semantic indexes and fewer in number compared to the keypoints extracted from other VPR approaches.

2) Model Comparison with the Baseline Text Spotting Utilized in the VPR Application. The TextPlace [22] model uses the pre-trained model of Textboxes++ [34] algorithm as the primary text extraction in their framework for the VPR application.

In this section, additional experiments to compare the proposed model with Textboxes++ [34] are conducted and provided quantitative and qualitative results to show how the proposed model performs for text instances that appear in the wild images using the benchmark dataset, ICDAR15 [25], as in [34].

Fig. 5: Comparing the (a) Text features used in the proposed E2E text detection model and (b) keypoint features used in different VPR techniques [11], [45].

Table 2 shows the quantitative comparison of the Textboxes++ that is used as a baseline [22] and the proposed model using the well-known text detection and end-to-end text spotting evaluation metrics [25]. As seen, the proposed approach outperformed the [34] in both detection and end-to-end spotting tasks. It achieves an H-mean detection performance of 86.5% compared to 82.9% in [34]. It also surpasses the Textboxes++ method with a large margin of $\sim 16\%$ in end-to-end F-measure performance. Furthermore, the proposed model is more suitable for real-time detection and recognition as it provides better FPS. These performances confirm the proposed models' good generalization and efficiency on challenging and unseen VPR dataset, SCTP (see Table 1).

To see how the proposed algorithm performs in challenging cases of the ICDAR15 dataset, a qualitative comparison of the proposed model with failure cases in [34] is provided. As shown in Fig.6, the proposed model successfully predicted most of the failure cases. Since text instances in the wild images usually appear with arbitrary shapes, it is important to use a model that better captures any shape of the scene text. The results in the last column in Fig. 6 also show that the proposed pipeline is capable of accurately outputting polygon bounding boxes for curved text instances, whereas Textboxes++ fails to detect.

3) Qualitative results on the challenging text sample images in the wild. To show the proposed model's capability and its limitation on challenging real-world scenarios, more qualitative results by showcasing visual examples of successful and unsuccessful predictions are provided. As shown, the proposed model performed well

on different challenging text instances in the wild images, such as partial occlusion, complex fonts, Illumination variation, oriented text, and curved text.

Fig. 6: Comparison between the proposed end-to-end model (bottom row) with Textboxes++ [34] algorithm (top row). The red boxes in the top row images show the failure cases (Images are taken from [34]), and the cyan text and boxes show our results. The orange arrows point to text regions where the proposed model could successfully predict failure text instances of Textboxes++.

All the images in Fig. 7 are selected from the datasets different from ICDAR15 used during fine-tuning and testing. For example, there is no curved annotation in ICDAR15, but the proposed model could accurately bound a curved bounding box around those text instances (Fig. 7.b, Fig. 7.c, and Fig. 7.d.). The performance and other successful predictions show that the proposed model could be generalizable on unseen sample text instances from the TotalText [76] and CTW1500 [77] datasets designed for irregular text detection and recognition and different challenges than ICDAR15. In addition, another experiment by applying partial occlusion on the characters is conducted. As seen in the letter 'G' in Fig. 7.b, letter 's' in Fig. 7.c, letter 'c,' 'k,' and letter 'r' in Fig. 7.d., the proposed model correctly recognizes those letters, confirming its capability of partial occlusion detection and recognition due to the capability of individual character spotting and masking the input images of the proposed architecture. However, from Fig. 7, the proposed model performed poorly on low-resolution and low-contrast text instances.

4) Inference Speed of the model. This work also experiments with the inference speed of the proposed model and compares it with [22] in terms of Frames Per Second (FPS). To that effect and for a fair comparison, an RTX 3080Ti GPU is used that has a similar memory used in [22] and presented in [34]. The proposed model outperformed the TextPlace method by a large margin, achieving a $\sim$ 11 FPS in compared to 2.3 FPS in [22].

G. Limitations and Future Work

As mentioned in the previous sections, although the proposed model performed well on many challenging cases in the wild images, there are many shortcomings that can be improved or addressed in future work. First, the model needed character annotation to be trained. These types of annotations are expensive to prepare. To address this problem, weakly supervised or unsupervised learning techniques can be applied to the model.

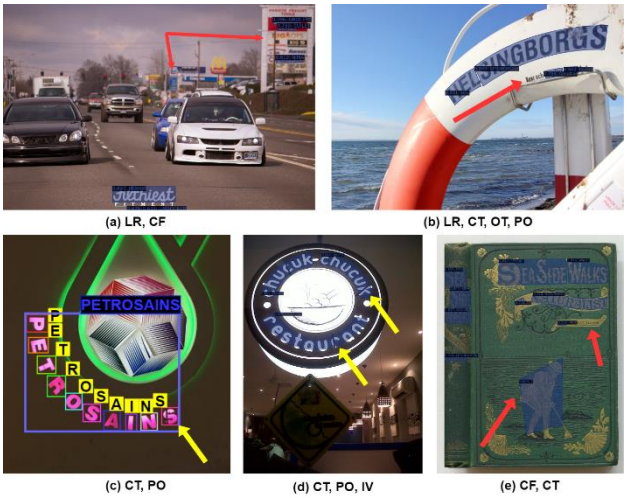


Fig. 7: Qualitative results of the proposed model on some challenging examples images, where PO: Partial Occlusion, CF: Complex Fonts, IV: Illumination Variation, LR: Low Resolution, OT: Oriented Text, and CT: Curved Text. The red and yellow arrows represent failure cases and the partially occluded characters, respectively. The above images are selected from two benchmarks: TotalText and CTW100 datasets designed for curved text detection and recognition in the wild. The proposed model needs to be trained in these images. The accurate detection of the text instances shows its generalizability on unseen images.

The proposed model performs poorly on low-resolution, blurry, and high-occluded text instances, as seen in Fig. 7. These challenges are still open in many SOTA text detection and recognition in the wild methods, and humans may need help reading these types of text instances. However, one way to address these problems is to benefit from the recent advancement in natural language processing algorithms like combining the pre-trained language model modules like Generative Pre-training Transformer (GPT) [78] and compositionality techniques [79] in the text detection and recognition framework to help the model to guess the uncaptured characters in the text.

Conclusion

In this work, an end-to-end scene text spotting model for the visual place recognition task is presented. The proposed model has leveraged a robust SOTA backbone of pre-trained MAE and a modified multi-task transformer detector. The quantitative and qualitative experimental results have shown that the proposed model outperforms SOTA models in VPR, which confirms the robustness of the proposed end-to-end scene text detection and recognition model. It obtained the best performance in

terms of precision-recall for the benchmark VPR application dataset, called SCTP. The proposed model outperformed the baseline text detection and recognition technique, textboxes++, used in TextPlace by a large margin regarding precision, recall, and H-mean. It also achieved competitive performance with many SOTA text detection and recognition techniques. The qualitative ablation experiments also confirmed that the proposed model could spot many challenging text instances in the wild images, including rotated and curved, complex fonts, partial illumination variation, and occlusion. The limitations and future work to improve the performance of the proposed model are also discussed. Other applications besides VPR include different facets of localization and mapping. Detecting and recognizing text allows the potential to leverage semantics and the features related to the detected text to localize better and map instead of just using indirect features.

Acknowledgment

We would like to thank the Ontario Centers of Excellence (OCE), the Natural Sciences and Engineering Research Council of Canada (NSERC), and ATS Automation Tooling Systems Inc., Cambridge, ON, Canada, for supporting this research work.

Author Contributions

Z. Raisi collected the data, implemented the code, carried out the analysis, and wrote paper. Dr. J. Zelek interpreted the results and supervised the research.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication or falsification, double publication and, or submission, and redundancy, have been completely witnessed by the authors.

Abbreviations

ViT	Vision Transformer
FPS	Frames per Second
SCTP	Self-Collected Text Place
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
DETR	Detection using Transformers
SOTA	State Of the Art
VPR	Visual Place Recognition
MAE	Masked Autoencoders
SLAM	Simultaneous Localization and Mapping

References

[1] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, L. Van Gool, "Night-to-day image translation for retrieval-based localization," in Proc. 2019 International Conference on Robotics and Automation (ICRA): 5958–5964, 2019.

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Proc. IEEE/CVF International Conference on Computer Vision: 5297–5307, 2016.

[3] R. Atienza, "Vision transformer for fast and efficient scene text recognition," Document Analysis and Recognition – ICDAR 2021. Springer International Publishing, pp. 319–334, 2021.

[4] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, H. Lee, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in Proc. International Conference on Computer Vision (ICCV), 2019.

[5] Y. Baek, B. Lee, D. Han, S. Yun, H. Lee, "Character region awareness for text detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[6] Y. Baek, S. Shin, J. Baek, S. Park, J. Lee, D. Nam, H. Lee, "Character region attention for text spotting," ArXiv, vol. abs/2007.09629, 2020.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers," arXiv preprint arXiv:2005.12872, 2020.

[8] W. Chan, C. Saharia, G. Hinton, M. Norouzi, N. Jaitly, "Imputer: Sequence modeling via imputation and dynamic programming," arXiv preprint arXiv:2002.08926, 2020.

[9] C. K. Ch'ng, C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in Proc. IAPR International Conference on Document Anal. and Recognition (ICDAR), 1: 935–942, 2017.

[10] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al., "Rethinking attention with performers," arXiv preprint arXiv:2009.14794, 2020.

[11] M. Cummins, P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," Int. J. Rob. Res., 27(6): 647–665, 2008.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[13] S. Fang, H. Xie, Y. Wang, Z. Mao, Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition: 7098–7107, 2021.

[14] W. Feng, W. He, F. Yin, X. Y. Zhang, C. L. Liu, "Textdragon: An end-to-end framework for arbitrarily shaped text spotting," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition: 9076–9085, 2019.

[15] S. Garg, T. Fischer, M. Milford, "Where is your place, visual place recognition?" arXiv preprint arXiv:2103.06443, 2021.

[16] A. Gupta, A. Vedaldi, A. Zisserman, "Synthetic data for text localization in natural images," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 2315–2324, 2016.

[17] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., "A survey on the visual transformer," arXiv preprint arXiv:2012.12556, 2020.

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, R. Girshick, "Masked autoencoders are scalable vision learners," arXiv preprint arXiv:2111.06377, 2021.

[19] K. He, G. Gkioxari, P. Dollar, R. Girshick, "Mask R-CNN, " in Proc. IEEE International Conference on Computer Vision: 2961–2969, 2017.

[20] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 770–778, 2015.

[21] S. Hochreiter, J. Schmidhuber, "Long short-term memory," Neural Comput., 9(8): 1735–1780, 1997.

[22] Z. Hong, Y. Petillot, D. Lane, Y. Miao, S. Wang, "Textplace: Visual place recognition and topological localization through reading scene texts," in Proc. IEEE/CVF International Conference on Computer Vision: 2861–2870, 2019.

[23] M. Iwamura, N. Morimoto, K. Tainaka, D. Bazazian, L. Gomez, D. Karatzas, "ICDAR2017 robust reading challenge on omnidirectional video," in Proc. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 1: 1448–1453, 2017.

[24] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition, " arXiv preprint arXiv:1406.2227, 2014.

[25] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al., "ICDAR 2015 competition on robust reading," in Proc. International Conference on Document Analysis and Recognition (ICDAR): 1156–1160, 2015.

[26] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. I Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. De Las Heras, "ICDAR 2013 robust reading competition," in Proc. International Conference on Document Analysis and Recognition: 1484–1493, 2013.

[27] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, "Transformers in vision: A survey," arXiv preprint arXiv:2101.01169, 2021.

[28] Y. Kittenplon, I. Lavi, S. Fogel, Y. Bar, R. Manmatha, P. Perona, "Towards weakly-supervised text spotting using a multi-task transformer," arXiv preprint arXiv:2202.05508, 2022.

[29] A. B. Laguna, K. Mikolajczyk, "Key. net: Keypoint detection by handcrafted and learned CNN filters revisited," IEEE Trans. Pattern Anal. Mach. Intell., 45(1): 698-711, 2022.

[30] J. Lee, S. Park, J. Baek, S. Joon Oh, S. Kim, H. Lee, "On recognizing texts of arbitrary shapes with 2D self-attention," in Proc. IEEE CVPR: 546–547, 2020.

[31] H. Li, P. Wang, C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in Proc. 2017 IEEE International Conference on Computer Vision (ICCV): 5248–5256, 2017.

[32] Y. Li, S. Xie, X. Chen, P. Dollar, K. He, R. Girshick, "Bench-marking detection transfer learning with vision transformers," arXiv preprint arXiv:2111.11429, 2021.

[33] M. Liao, G. Pang, J. Huang, T. Hassner, X. Bai, "Mask textspotter v3: Segmentation proposal network for robust scene text spotting," in Proc. Computer Vision–ECCV 2020: 16th European Conference, Part XI 16: 706–722, 2020.

[34] M. Liao, B. Shi, X. Bai, "Textboxes++: A single-shot oriented scene text detector," IEEE Trans. Image Process., 27(8): 3676–3690, 2018.

[35] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. L. Zitnick, "Microsoft coco: Common objects in context," in Proc Eur. Conference on Computer Vision. Springer: 740–755, 2014.

[36] V. Nazarzehi, R. Damani, "Decentralised optimal deployment of mobile underwater sensors for covering layers of the ocean," Indones. J. Electr. Eng. Comput. Sci., 25(2): 840–846, 2022.

[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, "SSD: Single shot multibox detector," in Proc. Eur. Conference on Computer Vision. Springer: 21–37, 2016.

[38] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, J. Yan, "FOTS: Fast oriented text spotting with a unified network," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 5676–5685, 2018.

[39] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition: 9809–9818, 2020.

[40] Y. Liu, C. Shen, L. Jin, T. He, P. Chen, C. Liu, H. Chen, "Abcnet v2: Adaptive bezier-curve network for real-time end-to-e end text spotting," arXiv preprint arXiv:2105.03620, 2021.

[41] S. Lowry, N. S. Underhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, M. J. Milford, "Visual place recognition: A survey," IEEE Trans. Rob., 32(1): 1–19, 2015.

[42] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, "ICDAR 2003 robust reading competitions," in Proc. Seventh Int. Conference on Document Analysis and Recognition: 682– 687, 2023.

[43] P. Lyu, M. Liao, C. Yao, W. Wu, X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in Proc. Eur. Conference on Computer Vision (ECCV) : 67–83, 2018.

[44] C. Masone, B. Caputo, "A survey on deep visual place recognition," IEEE Access, 9: 19516–19547, 2021.

[45] M. J. Milford, G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in Proc. IEEE International Conference on Robotics and Automation: 1643–1649, 2012.

[46] A. Mishra, K. Alahari, C. V. Jawahar, "Scene text recognition using higher order language priors," in BMVC, 2012.

[47] S. Qin, A. Bissacco, M. Raptis, Y. Fujii, Y. Xiao, "Towards unconstrained end-to-end text spotting," in Proc. IEEE/CVF International Conference on Computer Vision: 4704–4714, 2019.

[48] T. Q. Phan, P. Shivakumara, S. Tian, C. Lim Tan, "Recognizing text with perspective distortion in natural scenes," in Proc. IEEE International Conference on Computer Vision: 569–576, 2013.

[49] Z. Raisi, M. Naiel, P. Fieguth, S. Wardell, J. Zelek, "2d positional embedding-based transformer for scene text recognition," J. Comput. Vision Imaging Syst., 6(1): 1–4, 2021.

[50] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, J. Zelek, "Text detection and recognition in the wild: A review," arXiv preprint arXiv:2006.04305, 2020.

[51] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell, J. Zelek, "2lspe: 2d learnable sinusoidal positional encoding using a transformer for scene text recognition," in Proc. Conference on Robots and Vision (CRV): 119–126, 2021.

[52] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell, J. S. Zelek, "Transformer-based text detection in the wild," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops: 3162–3171, 2021.

[53] Z. Raisi, G. Younes, J. Zelek, "Arbitrary shape text detection using transformers," in Proc. IEEE International Conference on Pattern Recognition (ICPR): 3238-3245, 2022.

[54] Z. Raisi, J. Zelek, “Occluded text detection and recognition in the wild,” in IEEE Proceeding Conference on Robots and Vision (CRV): 140-150, 2022.

[55] Z. Raisi, J. S. Zelek, “End-to-end scene text spotting at character level,” J. Comput. Vision Imaging Syst., 7(1): 25-27, 2021.

[56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 779–788, 2016.

[57] S. Ren, K. He, R. Girshick, J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. in Neural Info. Process. Syst.: 91–99, 2015.

[58] A. Risnumawan, P. Shivakumara, C. S. Chan, C. L. Tan, “A robust arbitrary text detection system for natural scene images,” Expert Syst. Appl., 41(18): 8027–8048, 2014.

[59] D. E. Rumelhart, G. E. Hinton, R. J. Williams, “Learning representations by back-propagating errors,” Nature, 323(6088): 533–536, 1986.

[60] A. Shahab, F. Shafait, A. Dengel, “ICDAR 2011 robust reading competition challenge 2: Reading text in scene images,” in Proc. International Conference on Doc. Anal. and Recognition: 1491–1496, 2011.

[61] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, X. Bai, “Aster: An attentional scene text recognizer with flexible rectification,” IEEE Trans. Pattern Anal. Mach. Intell., 41(9): 2035-2048, 2018.

[62] Y. Sun, Z. Ni, C.-K. Chng, Y. Liu, C. Luo, C. C. Ng, J. Han, E. Ding, J. Liu, D. Karatzas, et al., “ICDAR 2019 competition on large-scale street view text with partial labeling –RRC-LSVT,” arXiv preprint arXiv:1909.07741, 2019.

[63] Y. Tay, M. Dehghani, D. Bahri, D. Metzler, “Efficient transform- ers: A survey,” arXiv preprint arXiv:2009.06732, 2020.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, “Attention is all you need,” in Proc. Advances in Neural Information Processing Systems (NIPS 2017): 5998– 6008, 2017.

[65] K. Wang, S. Belongie, “Word spotting in the wild,” in Proc. Eur. Conference on Computer Vision. Springer: 591–604, 2010.

[66] C. Yao, X. Bai, W. Liu, Y. Ma, Z. Tu, “Detecting texts of arbitrary orientations in natural images,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 1083–1090, 2012.

[67] L. Yuliang, J. Lianwen, Z. Shuaitao, Z. Sheng, “Detecting curve text in the wild: New dataset and new solution,” in arXiv preprint arXiv:1712.02170, 2017.

[68] X. Zhang, Y. Su, S. Tripathi, Z. Tu, “Text spotting transformers,” arXiv preprint arXiv:2204.01918, 2022.

[69] X. Zhang, L. Wang, Y. Su, “Visual place recognition: A survey from deep learning perspective,” Pattern Recognit., 113: 107760, 2021.

[70] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” arXiv preprint arXiv:2010.04159, 2020.

[71] S. X. Zhang, X. Zhu, J. B. Hou, C. Liu, C. Yang, H. Wang, X. C. Yin, “Deep relational reasoning graph network for arbitrary shape text detection,” in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 9699-9708, 2020.

[72] L. Xing, Z. Tian, W. Huang, M. R. Scott, “Convolutional character networks,” in Proc. the IEEE/CVF International Conference on Computer Vision: 9126-9136, 2019.

[73] I. Loshchilov, F. Hutter, “Decoupled weight decay regularization,” in Proc. International Conference on Learning Representations, 2018.

[74] G. Liao, Z. Zhu, Y. Bai, T. Liu, Z. Xie, “PSENet-based efficient scene text detection,” EURASIP J. Adv. Signal Process., 97(1), 1-13, 2021.

[75] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou,W. He, J. Liang, “East: an efficient and accurate scene text detector,” in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 5551-5560, 2017.

[76] C. K. Ch'ng, C. S. Chan, “TotalText: A comprehensive dataset for scene text detection and recognition,” in Proc. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 1: 935-942, 2017.

[77] L. Yuliang, J. Lianwen, Z. Shuaitao, Z. Sheng, “Detecting curve text in the wild: New dataset and new solution,” in arXiv preprint arXiv:1712.02170, 2017.

[78] D. M. Katz, M. J. Bommarito, S. Gao, P. Arredondo, Gpt-4 passes the bar exam. Available at SSRN 4389233.

[79] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, “Building machines that learn and think like people,” Behav. Brain sci., 40, 2017.

Biographies

Zobeir Raisi was born in Chabahar, Iran in 1987. He received his Ph.D. degree in 2022 from the Vision Image Processing Lab (VIPLab) at the Systems Design Engineering Department, University of Waterloo, Waterloo, Ontario, Canada. Currently, he is an assistant professor in the Department of Electrical Engineering at Chabahar Maritime University, Iran. His research interests include computer vision, artificial intelligence, robotics, and image processing.

- Email: zraisi@uwaterloo.ca
- ORCID: 0000-0002-1591-4492
- Web of Science Researcher ID: GLV-1410-2022
- Scopus Author ID: 54897975500
- Homepage: https://uwaterloo.ca/scholar/zraisi

John Zelek received his Ph.D. degree in Philosophy of Electrical Engineering from the Centre for Intelligent Machines (CIM), McGill University, Montreal, QC, Canada, in 1996. He is currently a Professor of the Systems Design Engineering Department, University of Waterloo, Waterloo, ON, Canada, and the Co-Director of the Vision Image Processing (VIP) Laboratory, University of Waterloo, ON, Canada. He has published over 300 refereed articles, has been a co-founder of five different startup companies from the University of Waterloo, and has been an advisor for various other companies. His research interests include computer vision, AI, robotics, infrastructure monitoring, autonomous vehicles, image processing, augmented reality, and assistive technology, to name a few.

- Email: jzelek@uwaterloo.ca
- ORCID: 0000-0002-8138-3546
- Web of Science Researcher ID: NA
- Scopus Author ID: 6603746225

Homepage: https://uwaterloo.ca/systems-design-engineering/profile/jzelek

How to cite this paper:
Z. Raisi, J. Zelek, "Text detection and recognition for robot localization" J. Electr. Comput. Eng. Innovations, 12(1): 163-174, 2024.
DOI: 10.22061/jecei.2023.9857.658
URL: https://jecei.sru.ac.ir/article_1986.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Merged LNA-Mixer with Wide Variable Conversion Gain and Low Noise Figure for WLAN Direct-Conversion Receivers

A. Bijari [*]***, M. A. Mallaki***

Deptartment of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 08 June 2023
Reviewed 07 August 2023
Revised 29 September 2023
Accepted 08 October 2023

Keywords:
Variable conversion gain
Low noise transconductance amplifier (LNTA)
Active mixer
Noise figure

[*]Corresponding Author's Email Address: *a.bijari@birjand.ac.ir*

Abstract

Background and Objectives: In wireless communications, receivers play an essential role. Among receiver architectures, the direct-conversion receiver (DCR) architecture has been selected due to its high level of integration and low cost. However, it suffers from DC offset due to self-mixing, I/Q imbalance, and flicker noise.

Methods: This paper presents a new LNA-mixer with variable conversion gain (VG-LM) for wireless local area network (WLAN) applications. A low noise transconductance amplifier (LNTA) is used as the transconductance stage in the Gilbert cell mixer. The wide variable conversion gain range is achieved by the change in LNTA's transconductance and transconductance of the mixer switching transistors.

Results: The proposed LNA-mixer is designed and simulated using 0.18µm CMOS technology in Cadence Spectre RF. The post-layout simulations exhibit the proposed circuit operates at 2.4 GHz with a bandwidth of 10 MHz. In addition, the conversion gain is changed from -3.9 dB to 23.9 dB with the variation of the controlled DC voltage from 0.5 to 1.8. At the high gain, the double-sideband noise figure (DSB-NF) is less than 3.7 dB, and its third-order intermodulation point (IIP3) is -9 dBm. The power consumption is 22 mW from the supply voltage of 1.8 V. The circuit occupies 743 µm×775 µm of core chip area.

Conclusion: Using the proposed circuit, the RF front end receiver does not need the low noise amplifier (LNA) and variable gain amplifier (VGA).

Introduction

In recent years, wireless communications have considerably developed, and the designers have sought to address the RF front-ends with higher level of integration and lower cost [1]-[7]. A direct conversion receiver (DCR) front-end suffers from DC offset due to self-mixing, I/Q imbalance, and flicker noise. However, it still is used widely due to small size, low cost, low power consumption, and the fewer number of external components [8], [9]. As illustrated in Fig. 1, the DCR front-end consists of RF bandpass filter (BPF), low-noise amplifier (LNA), I/Q mixer, variable gain amplifier (VGA), and lowpass filter (LPF).

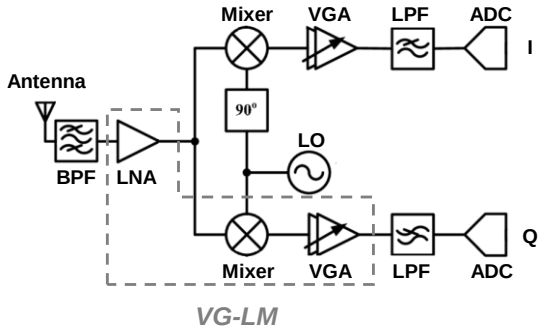


Fig. 1: Direct conversion front-end architecture.

As seen, the variable gain amplifier (VGA) is inserted between the mixer and the LPF to compensate the low conversion gain of the mixer and control the signal power level. Consequently, the VGA increases the dynamic range of the DCR [10]. The VGA is typically used along with the auxiliary circuits, such as common-mode feedback (CMFB), DC offset voltage cancellation, and the exponential current generator. The circuits provide a constant signal amplitude for the analog to digital converter (ADC) and eliminate the undesirable DC offset voltage [11].

Merging some blocks in RF front-end is an effective method to achieve high integration level and low cost. Martines et al. [12] designed a combined LNA and mixer with low power consumption and chip area. However, the proposed LNA-mixer exhibits low conversion gain and high noise figure (NF). Ryu et al. [13] proposed a variable gain mixer with automatic gain control (AGC). This proposed design achieves a wide variable conversion gain range with a maximum conversion gain of 10 dB. In addition, the AGC circuit suffers from the design complexity to obtain the dB linear function with MOS transistors. Kolios and Kalivas [14] used the dynamic current bleeding technique to realize the variable-conversion gain mixer (VCG-mixer). Moreover, they applied the inter-stage inductors to cancel the parasitic capacitances seen at the source of switching transistors. However, the circuit presents a low conversion gain range and occupies a large chip area. Wang and Saavedra [15] employed a Gilbert-Cell mixer with the reconfigurable gate widths in the RF transconductance transistor, thereby varying the conversion gain. However, the mixer does not obtain a continuous conversion gain range. Wu and Chou [16] applied forward body bias control to a body terminal of the RF transconductance and LO switching stages to achieve a mixer with variable conversion gain. The proposed mixer consumes low power, but it suffers from low conversion gain and a narrow tunable range. Kalamani [17] employed the double balanced mixer and the differential LNA to achieve high conversion gain, but it consumes high power. Hu et al. [18] offered a LNA-mixer with improved conversion gain and NF. The gain boosted and the current bleeding techniques are exploited to achieve a high gain and an appropriate NF. Gladson et.al. [19] used a low-noise transconductance amplifier (LNTA) in the RF transconductance stage of the down conversion mixer to reduce the noise of the switching stage. Moreover, the linearity of the low-noise stage is improved by using post-distortion based harmonic cancellation technique, that provides an enhanced spurios free dynamic range (SFDR) of up to 81.88 dB. Cao et al. [20] proposed a digitally controlled dedicated short range communications (DSRC) receiver operating at 5.8 GHz, specifically designed for the Chinese electronic toll collection system. The design utilizes a digital baseband for controlling LNA and mixer circuits. However, the LNA cannot exhibit a continuously variable gain, and it is characterized by four discrete modes. Guo et al. [21] employed a highly linear wideband differential LNTA for current-mode SAW-less receiver architecture. The design is developed by using an active-combiner feedback and complementary multi-gated transistor (MGTR) configurations, that provides high linearity with simultaneous compensations for the second- and third-order nonlinearity of transistors.

In this paper, the structural innovation is done to merge the LNA with down-conversion mixer, and VGA. The proposed variable gain LNA-Mixer (VG-LM) can control the conversion gain, translate the RF input to IF output with low noise performance, simultaneously, and exhibits higher level of integration. In the proposed circuit, an LNTA is inserted in the RF transconductance stage to control the conversion gain continuously by up to 20 dB. In addition, the proposed LNTA enhances LNA-mixer conversion gain (CG), thereby improving the noise performance effectively. The proposed VG-LM is designed over the RF frequency of 2.4 GHz for WLAN applications. The remainder of the paper is organized as follows. Section 2 presents the VG-LM structure and details the proposed LNTA. In Section 3, the simulation results are discussed, and the performance of the VG-LM is compared with previous studies. Finally, the conclusion is presented in Section 4.

Design of Proposed VG-LM

A. Conventional Gilbert Cell Mixer

Fig. 2 illustrates the double-balanced active mixer based on the Gilbert cell. As illustrated, it consists of the RF input, current switch, and load sections. The RF input is called the RF transconductance, which converts a voltage to a current signal by transistors of MRF. The RF signal passes from the LO switching for frequency translation, and it is commutated by transistors of MSW. Finally, the IF current converts to IF voltage by load resistors [22], [23].

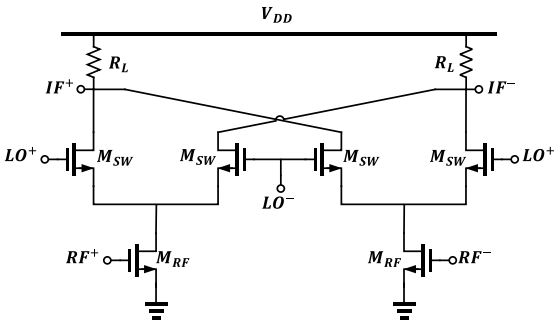


Fig. 2: Schematic of the conventional downconversion active mixer.

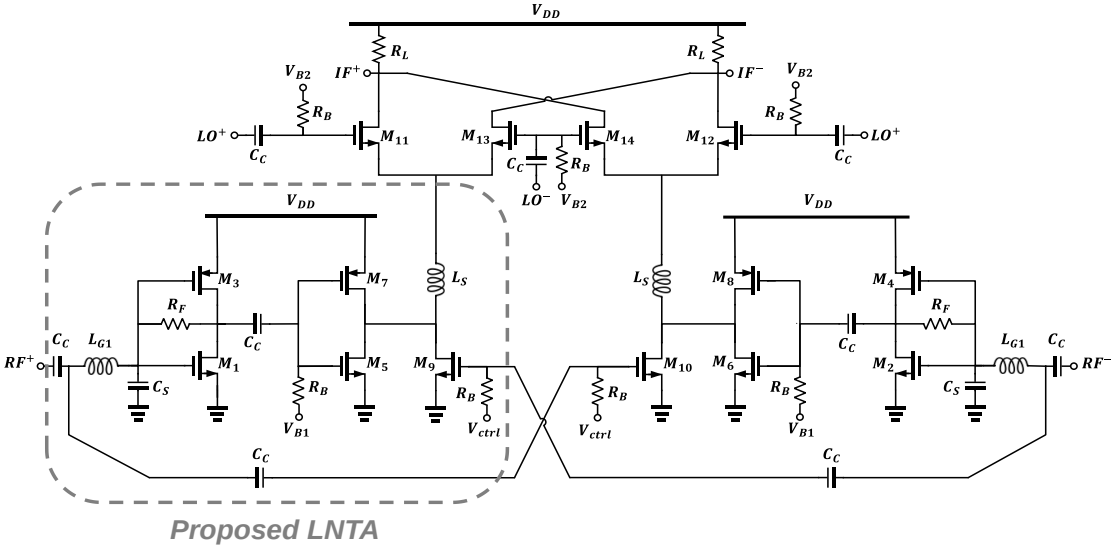


Fig. 3: Schematic of the proposed VG-LM.

The voltage conversion gain (CG) of the mixer is reduced by gradual LO transitions and the parasitic capacitor (C_P) that is seen at the source of switching transistors [17]. Therefore, the CG of the active mixer is defined as follows:

$$CG = \frac{2}{\pi} g_{m,RF} R_L \frac{g_{m,SW}(1-\alpha)}{\sqrt{C_P^2\,\omega^2 + g_{m,SW}^2}} \quad (1)$$

where $g_{m,RF}$ and $g_{m,SW}$ are the transconductances of the RF input and LO switching, respectively. α is equal to $2\Delta T/T_{LO}$, where ΔT is a fraction of each half cycle of the LO period (T_{LO}), that the LO transistors act as a balanced differential pair.

B. Proposed VG-LM

Fig. 3 illustrates the schematic of the proposed VG-LM. The proposed circuit exploits a low noise transconductance amplifier (LNTA) in the RF input stage of the mixer. The variable conversion gain can be realized by changing the LNTA's transconductance ($G_{m,LNTA}$) and transconductance of the mixer switching transistors through the DC control voltage (V_{ctrl}).

As shown in Fig. 4, the proposed LNTA consists of cascaded stages of a resistive shunt–shunt feedback amplifier (M_1 and M_3) and an inverter-based amplifier (M_5 and M_7). When the controllable bias voltage (V_{ctrl}) is low enough, M_9 operates in the saturation region, and the input signal is amplified with the common-source amplifier. Since the DC current of M_3 is reused by M_1, the power consumption of the input stage is reduced. This stage is self-biased by the feedback resistor of R_F and coupled to the second stage by the large *ac* coupling capacitance of C_C. The input stage of LNTA is designed to achieve the high voltage gain and low noise figure (NF), while the second stage provides linear-in-dB gain control characteristics. The inverter-based amplifier is loaded by M_9 with controllable bias voltage.

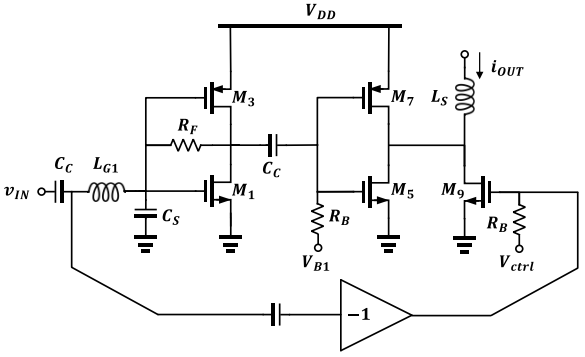


Fig. 4: Schematic of the proposed LNTA.

The current passing through M_9 can control the effective transconductance of the second stage. The transconductance of the LNTA can be degraded by the parasitic capacitances seen at the source of mixer switching transistors at high frequency. Therefore, an inductive series peaking L_S is inserted at the output.

C. Input Matching

The *LC* (L_{G1} and C_S) and resistive feedback techniques are inserted in the proposed input matching network. The series inductor, L_{G1}, is used to extend the RF bandwidth and reduce the noise contributed by LNTA. Assuming that the total impedance due to the gate-drain capacitances of M_1 and M_3 (C_{gd1}+C_{gd3}) is relatively higher than R_F, the small-signal equivalent circuit of the proposed LNTA is simplified, as illustrated in Fig. 5. As illustrated, the capacitor of C_S and the gate-source capacitances of M_1 and M_3 are merged as C_{eq1}. The total of the gate-source capacitances of M_5 and M_7 and their input Miller capacitances is represented by C_{eq2}, and C_{eq3} is the parasitic capacitance seen at the drain of M_9.

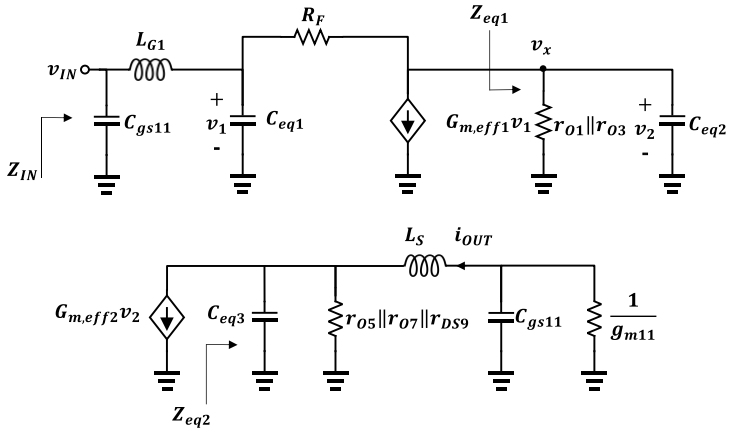


Fig. 5: Small-signal equivalent circuit of the proposed LNTA.

By neglecting the gate-drain capacitances, C_{eq3} equals C_{db5}+C_{db7}+C_{db9}. The input impedance is equal to the series combination of L_{G1} with the parallel combination of C_{eq1} and the impedance seen at the gate of M_1 and M_2 and is given by:

$$Z_{IN} = \frac{1}{C_{gs11}s} \| \left(L_{G1}s + \left(\frac{1}{C_{eq1}s} \| \frac{R_F + Z_{eq1}}{1 + G_{m,eff1}Z_{eq1}} \right) \right) \tag{2}$$

where $G_{m,eff1}$=g_{m1}+g_{m3} is the effective transconductance of the input stage, and Z_{eq1} represents the load impedance and is given by:

$$Z_{eq1} = (r_{O1}\|r_{O3}) \| \frac{1}{C_{eq2}s} \approx (r_{O1}\|r_{O3}) \tag{3}$$

where r_O represents the MOS output resistance. By substituting (3) in (2) and neglecting channel-length modulation, Z_{IN} can be rewritten as:

$$Z_{IN} = \frac{1}{C_{gs11}s} \| \left(L_{G1}s + \left(\frac{1}{C_{eq1}s} \| \frac{1}{G_{m,eff1}} \right) \right) \tag{4}$$

By assuming that C_S is high enough to satisfy $G_{m,eff1}/C_{eq1}$<<1/√($L_{G1}C_{eq1}$), the (4) reveals there is a dominant pole near dc and a resonant zero close to ω_0=1/√($L_{G1}C_{eq1}$). At the frequency of ω_0=2π×2.4 rad/s, Z_{IN} is obtained as follows:

$$Z_{IN} = \frac{L_{G1}G_{m,eff1}}{C_{eq1}} = (L_{G1}\omega_0)^2 G_{m,eff1} \tag{5}$$

According to (5), the input matching can be achieved by proper choice of $G_{m,eff1}$, and L_{G1}. Thus, $G_{m,eff1,}$ and L_{G1} values are chosen equal to 30 mA/V and 2.7 nH, respectively, to achieve input matching at 2.4 GHz.

D. Gain AnalysisSection Headings

Based on the small-signal equivalent circuit illustrated in Fig. 5 and (1), the CG of the proposed mixer can be obtained as follows:

$$CG = \frac{2}{\pi} G_{m,LNTA} R_L \frac{g_{m11}(1-\alpha)}{\sqrt{C_{gs11}^2\,\omega^2 + g_{m11}^2}} \cong \frac{2}{\pi} G_{m,LNTA} R_L (1-\alpha) \tag{6}$$

where $G_{m,LNTA}$ represents the total transconductance of LNTA and is given by:

$$G_{m,\,LNTA} = \frac{i_{OUT}}{v_{IN}} = \frac{i_{OUT}}{v_x} \times \frac{v_x}{v_{IN}} \tag{7}$$

where,

$$\frac{v_x}{v_{IN}} = \frac{-(G_{m,eff1}R_F - 1)Z_{eq1}}{C_{eq1}L_{G1}Z_{eq1x}s^2 + L_{G1}G_{m,eff1}Z_{eq1}s + Z_{eq1x}} \tag{8}$$

and,

$$\frac{i_{OUT}}{v_x} = \frac{G_{m,eff2}g_{m11}Z_{eq2}}{1 + g_{m11}Z_{eq2}} \times \frac{1}{1 + L_{G2}C_{eq2}s^2} \tag{9}$$

where Z_{eq1x}=Z_{eq1}+R_F, and $G_{m,eff2}$=g_{m5}+g_{m7} is the effective transconductance of the second stage, and Z_{eq2} represents the load impedances, and it is expressed as:

$$Z_{eq2} = \left(r_{O5}\|r_{O7}\|r_{DS9}\| \frac{1}{C_{eq3}s} \right) \| \left(L_S s + \left(\frac{1}{C_{gs11}s} \| \frac{1}{g_{m11}} \right) \right) \tag{10}$$

The channel resistance of M_9, r_{DS9}, depends on the controllable bias voltage of V_{ctrl}, and it can vary from r_{O9} to R_{ON9}=1/($\mu_n C_{OX}(W/L)(V_{ctrl}$-$V_{th})$) when M_9 is driven into the triode region. The π-network consisting of C_{eq3}, L_S, and C_{gs11} presents an infinite impedance at the frequency of 1/√($C_{eq1}C_{gs11}L_S$/(C_{eq1}+C_{gs11})), and therefore Z_{eq2} is simplified as:

$$Z_{eq2} = \left(r_{O5}\|r_{O7}\|r_{DS9}\| \frac{1}{g_{m11}} \right) \tag{11}$$

According to (11) and neglecting channel-length modulation, the (9) is approximately simplified as follows

at the resonance frequency of $1/\sqrt{(L_{G2}C_{eq2})}$:

$$\frac{i_{OUT}}{v_x} = \frac{G_{m,eff2}g_{m11}\left(r_{DS9}\|\frac{1}{g_{m11}}\right)}{\sqrt{2}\left(1+g_{m11}\left(r_{DS9}\|\frac{1}{g_{m11}}\right)\right)} \quad (12)$$

In addition, neglecting channel-length modulation and according to (3), the (8) is rewritten as:

$$\frac{v_x}{v_{IN}} = -\frac{G_{m,eff1}R_F}{C_{eq1}L_{G1}s^2 + L_{G1}G_{m,eff1}s + 1} \quad (13)$$

By proper choosing of L_{G1} and resonating with C_{eq1}, the gain peaking can be achieved, and (13) can be simplified at $\omega_0=1/\sqrt{(L_{G1}C_{eq1})}$ as follows:

$$\frac{v_x}{v_{IN}} = -\frac{R_F}{\omega_0 L_{G1}} \quad (14)$$

Consequently, by replacing (12) and (14) in (7), $G_{m,LNTA}$ is calculated as follows:

$$G_{m,\,LNTA} = \frac{i_{OUT}}{v_{IN}} \cong -\frac{G_{m,eff2}R_F g_{m11}\left(r_{DS9}\|\frac{1}{g_{m11}}\right)}{\sqrt{2}L_{G1}\omega_0\left(1+g_{m11}\left(r_{DS9}\|\frac{1}{g_{m11}}\right)\right)} \quad (15)$$

when V_{ctrl} is increased from V_{th} to supply voltage, M_9 enters from saturation to triode region, and v_{DS9} is decreased. Thus, r_{DS9} is reduced as well as $1/g_{m11}$. Moreover, the bias current of M_9 is effectively enhanced by increasing V_{ctrl}, and therefore, $G_{m,eff2}$ is reduced. Fig. 6 illustrates the g_{m11} and $G_{m,LNTA}$ versus V_{ctrl} under the input matching condition of Z_{IN}=50 Ω and operating M_5 in the triode region.

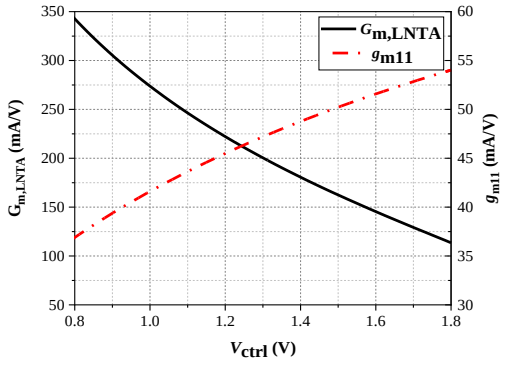


Fig. 6: Theoretical $G_{m,LNTA}$, and g_{m11} versus controllable bias voltage of M_9.

As can be seen, the conversion gain exhibits a wider range of variations than $G_{m,LNTA}$ due to increasing g_{m11} with V_{ctrl}.A primary section heading is enumerated by a Roman numeral followed by a period and is centered above the text. A primary heading should be in capital letters.

E. Noise Analysis

The thermal noise of all transistors and resistors and the flicker noise of the switching transistors are considered the noise sources of the downconversion active mixer. However, the flicker noise of the RF input section is translated up by ω_{LO}, and it does not appear at the baseband [22]. Therefore, the flicker noise contributed by LNTA is not considered in the noise analysis of the proposed VG-LM. The output thermal noise of the conventional downconversion active mixer is expressed as [17]:

$$\overline{v_{n.OUT}^2} = \overline{I_{n.M_{RF}}^2}R_L^2(1-\alpha) + \overline{v_{n.M_{SW}}^2}\left(2\,g_{m,SW}^2\alpha + C_P^2\omega^2(1-\alpha)\right)R_L^2 + 8kTR_L \quad (16)$$

where,

$$\overline{I_{n.M_{RF}}^2} = 4kT\gamma g_{m,RF} \quad (17)$$

$$\overline{v_{n.M_{SW}}^2} = \frac{4kT\gamma}{g_{m,SW}} \quad (18)$$

where k and T are Boltzmann constant, and the absolute temperature, respectively, and γ represents the MOS transistor thermal noise coefficient. For the proposed VG-LM, the first term in (16) should be substituted by the output thermal noise current contributed by LNTA ($I^2_{n,LNTA}$) as follows:

$$\overline{v_{n.OUT}^2} = \overline{I_{n.LNTA}^2}R_L^2(1-\alpha) + \overline{v_{n.M_{SW}}^2}\left(2\,g_{m,SW}^2\alpha + C_P^2\omega^2(1-\alpha)\right)R_L^2 + 8kTR_L \quad (19)$$

The simplified model for LNTA noise analysis is derived as illustrated in Fig. 7. Based on Fig. 7 and neglecting the channel-length modulation, $I^2_{n,LNTA}$ is calculated as:

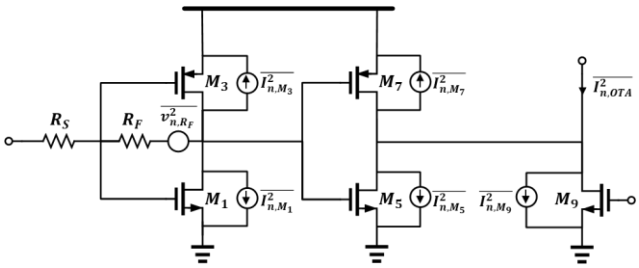


Fig. 7: Simplified equivalent circuit of LNTA for noise calculation.

$$\overline{I_{n.LNTA}^2} = \left(\left(\overline{v_{n.R_F}^2} + \left(\overline{I_{n.M_1}^2} + \overline{I_{n.M_3}^2}\right)\left(\frac{R_F + R_S}{1 + G_{m,eff1}R_S}\right)^2\right)G_{m,eff2}^2 + \left(\overline{I_{n.M_5}^2} + \overline{I_{n.M_7}^2}\right)g_{m11}^2\left(r_{DS9} \| \frac{1}{g_{m11}}\right)^2 + \overline{I_{n.M_9}^2}\right) \quad (20)$$

Under perfect matching condition and operating M_9 in triode region, (20) is rewritten as:

$$\overline{I_{n.LNTA}^2} = 4kT\left(\left(R_F + \gamma G_{m,eff1}\frac{(R_F + R_S)^2}{4}\right)G_{m,eff2}^2 + \gamma G_{m,eff2} + \frac{1}{r_{DS9}}\right) \quad (21)$$

where r_{DS9} represents the channel resistance of M_9 when it enters the triode region. γ is the excess noise coefficient and its value is assumed 1.3 for 180 nm transistors [24]. By assuming R_F>>R_S, the single side-band noise figure contributed by LNTA ($NF_{SSB,LNTA}$) can be obtained as:

$$NF_{SSB,LNTA} \cong \frac{\pi^2\left(4 + \gamma G_{m,eff1}R_F\right)}{4G_{m,eff1}R_F(1-\alpha)} \times \frac{\pi^2\gamma}{G_{m,eff1}G_{m,eff2}R_F^2(1-\alpha)} \quad (22)$$

To theoretically evaluate the noise performance of the proposed VG-LM, Fig. 8 illustrates the double side-band NF contributed by LNTA. As can be seen, the NF is related to the effect of M_9, and increasing V_{ctrl} allows further increase of NF.

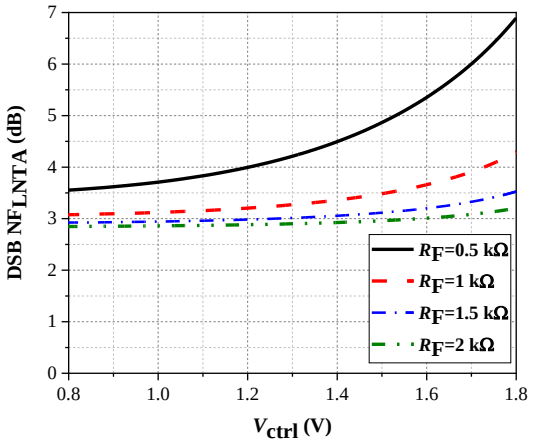


Fig. 8: Theoretical DSB NFLNTA versus controllable bias voltage of M_9 with α=0.1 and γ=1.33.

Moreover, for higher values of R_F, NF_{LNTA} exhibits less changes with increasing V_{ctrl}. However, choosing high values of R_F can degrade the input matching impedance.

Thus, the R_F value is chosen equal to 1.5 kΩ to achieve desirable input matching and noise performance along with high conversion gain. According to Fig. 6 and Fig. 8, the conversion gain of the proposed mixer can be continuously controlled by the bias voltage of M_9, while DSB NF is enhanced by less than 1 dB. However, as illustrated in Fig. 6, the transconductance of switching transistors (g_{m11}) is increased by increasing V_{ctrl}, and as a result, can enhance the flicker noise contributed by switching transistors.

Simulation Results and Discussion

The proposed VG-LM with an RF frequency of 2.4 GHz and IF frequency of 10 MHz is designed and simulated in 180 nm RF-TSMC CMOS process by Cadence Spectre RF. To reduce the effect of parasitic capacitors, the length of transistors is chosen to be 0.18 µm. Moreover, the width of transistors is designed and optimized by considering the above analysis and power consumption. Table 1 presents the optimized values of device sizes for the proposed VG-LM.

Table 1:. Device sizes of the proposed VG-LM

Device	Parameters	Value
Transistor (*W*/*L*)	$M_{1,2}$	(66 µm/0.18 µm)
	$M_{3,4}$	(32 µm/0.18 µm)
	$M_{5,6}$	(215 µm/0.18 µm)
	$M_{7,8}$	(15 µm/0.18 µm)
	$M_{9,10}$	(42 µm/0.18 µm)
	$M_{11\text{-}14}$	(408 µm/0.18 µm)
Inductor (nH)	L_{G1}	2.2
	L_S	5.9
Resistor (kΩ)	R_F	1.5
	R_B	20
	R_L	0.35
Capacitor (pF)	C_S	0.6
	C_C	5
Bias voltage (V)	V_{B1}	0.62
	V_{B2}	0.56
	V_{DD}	1.8

The VG-LM consumes 12.46 mA and 14.7 mA from a 1.8 V supply voltage when the conversion gain is maximum and minimum, respectively. The circuit layout of the proposed VG-LM is illustrated in Fig. 9. The chip area is about 0.57 mm^2 (743 µm×775 µm), including all the pads and guard rings. Fig. 10 shows the post-layout simulated input return loss. By changing the control bias voltage of M_9 (V_{ctrl}), all the simulated S_{11} are less than −10 dB over the input frequency of 2.4 GHz. This performance proves an appropriate design of input matching and demonstrates the input matching is hardly affected by changing the V_{Ctrl}.

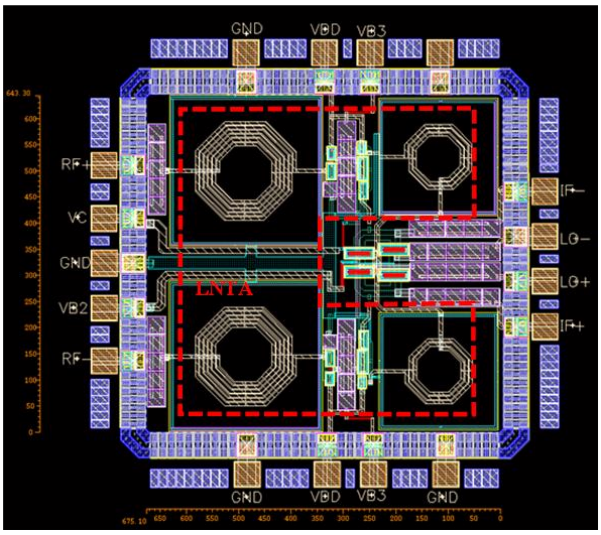


Fig. 9: Circuit layout of the proposed VG-LM.

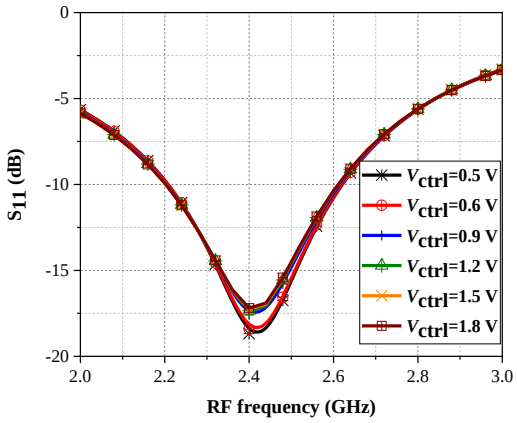
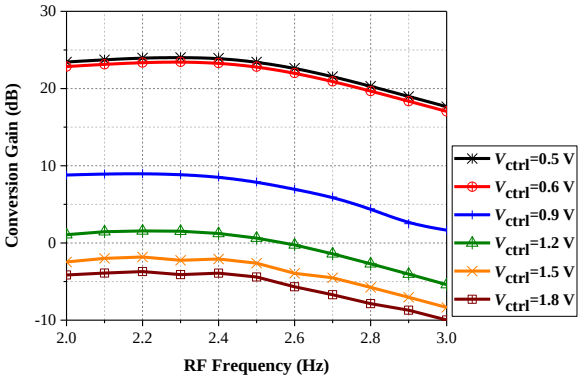


Fig. 10: Simulated results of input return loss with different values of V_{ctrl}.

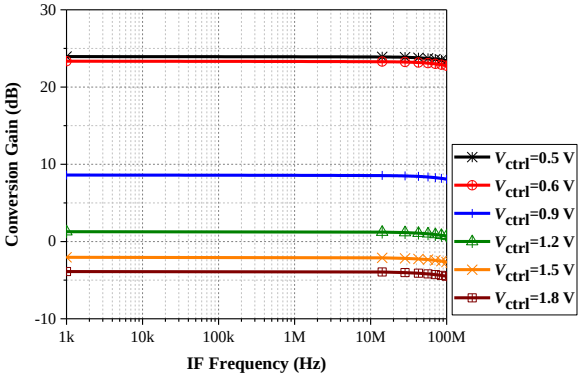
By considering P_{LO}= 0 dBm, the simulated conversion gain (CG) is illustrated in Fig. 11. As illustrated, the simulated conversion gain can be controlled continuously from 23.9 dB to -3.9 dB when V_{ctrl} is varied from 0.5 to 1.8 V at 2.4 GHz.

Fig. 12 presents the noise performance of the VG-LM in the intermediate frequency (IF) range of 1 kHz–10 MHz. At the IF frequency of 10 MHz, the simulated result indicates the DSB NF ranges from 3.74 dB to 6.71 dB when V_{ctrl} is changed from 0.5 V to 1.8. Moreover, it can be seen that DSB NF increases slightly as the conversion gain drops. As mentioned in the previous section, the flicker noise is relatively affected by V_{ctrl} changes.

To evaluate the nonlinear performance, the two tones with 4.125 MHz spacing are applied to the proposed VG-LM. Fig. 13 illustrates simulated input third-order intercept point (IIP3) at V_{ctrl}=0.5 V and V_{ctrl}=1.8 V. As illustrated, the VG-LM has IIP3 of -9 dBm and -6 dBm when the control bias voltage is 0.5 V and 1.8 V, respectively.



(a)



(b)

Fig. 11: Simulation results of the conversion gain versus (a) RF frequency and (b) IF frequency.

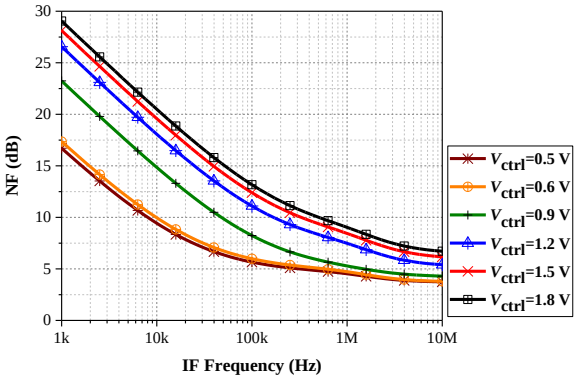


Fig. 12: Simulation results of NF versus IF frequency.

Furthermore, the output third-order intercept point (OIP3) values at the control voltages of 0.5 V and 1.8 V are equal to 17.5 dBm and 7.8 dBm, respectively. In addition, to evaluate the limitations arising from both noise and interference, the spurious-free dynamic range (SFDR) is evaluated as follows [22]:

$$SFDR\ [dB] = \frac{2}{3}(P_{IIP3}[dBm] + 174[dBm] - NF[dB] - 10\log(BW)) - SNR_{min} \quad (23)$$

By assuming minimum output SNR (SNRmin) of 10 dB and the bandwidth (BW) of 10 MHz, the resulting SFDR is equal to 50.84 dB at the control voltage of 0.5 V.

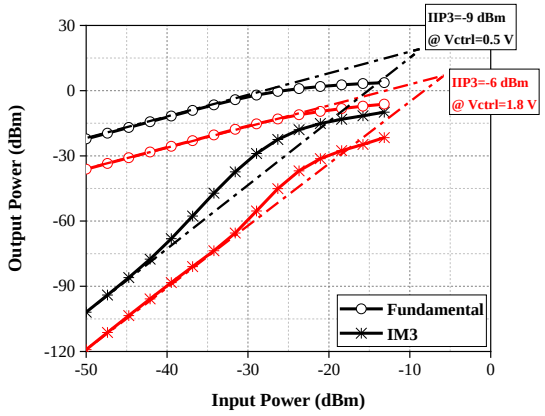


Fig. 13: Simulated IIP3 of the proposed VG-LM.

In Fig. 14, the simulated 1 dB-compression point (P_{1dB}) is illustrated for the maximum and minimum conversion gain. The results show that the VG-LM has a P_{1dB} of -21 dBm and -16 dBm at V_{ctrl}=0.5 V and 1.8 V, respectively.

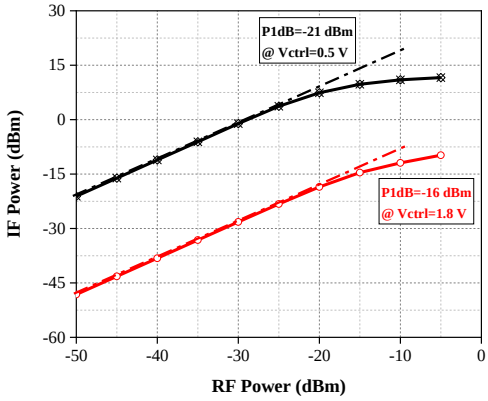
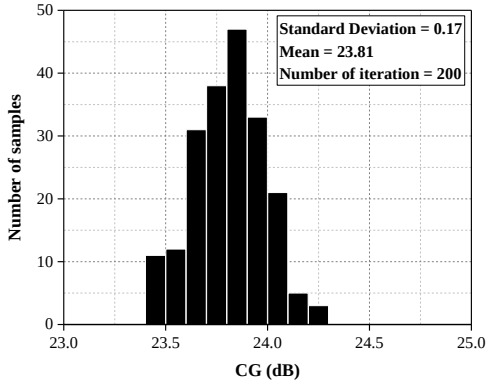
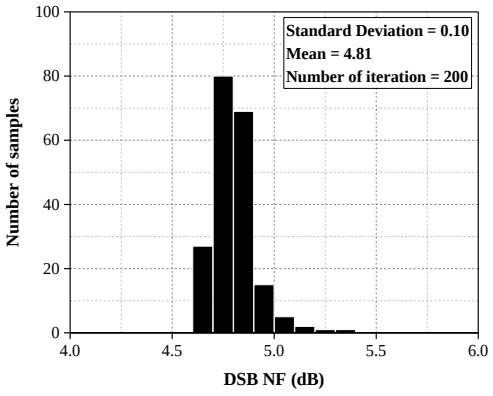


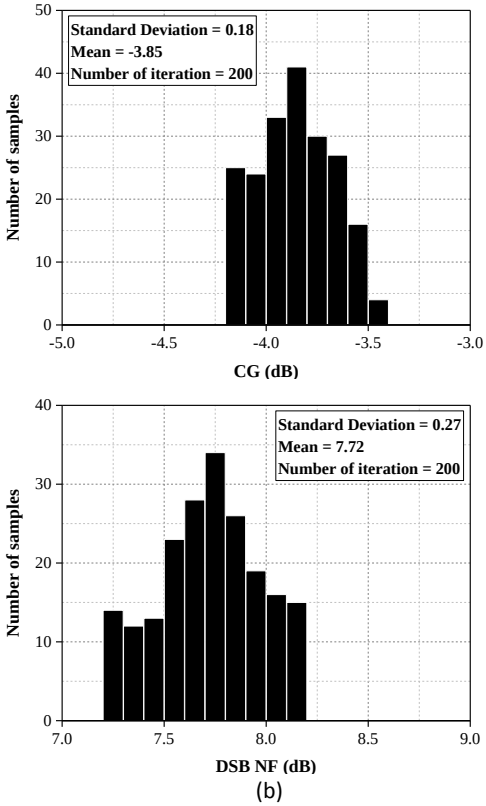
Fig. 14: Simulated P1dB of the proposed VG-LM.

The Monte Carlo results, including process variations and mismatch, are illustrated in Fig. 15 for 200 samples at the RF frequency of 2.4 GHz and the IF frequency of 10 MHz. The results show a mean CG of 23.81 dB with a standard deviation of 0.17 dB and DSB NF of 4.81 dB with a standard deviation of 0.1 dB at the maximum conversion gain. Moreovere, the mean CG is -3.85 dB with a standard deviation of 0.18 dB, and DSB NF is 7.72 dB with a standard deviation of 0.27 dB at the minimum conversion gain.





(a)



(b)

Fig. 15. Monte Carlo simulation results of 200 runs. (a) At the maximum CG. (b) At the minimum CG.

Consequently, the proposed mixer exhibits good stability against process variations and mismatch.

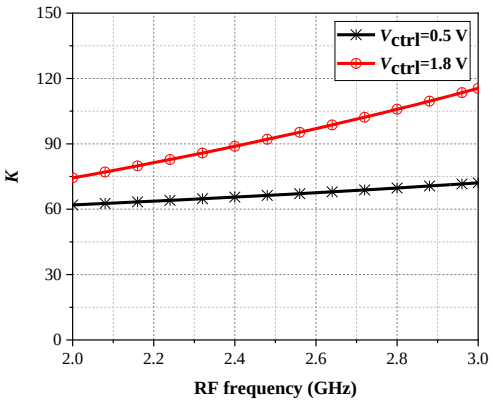
Fig. 16 (a), and (b) illustrates the stability factors based on the Sparameters to consider the stability of the proposed VG-LM. The necessary and sufficient conditions for unconditional stability are given as follows:

$$K = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + \Delta^2}{2|S_{12}||S_{21}|} > 1 \quad (24)$$

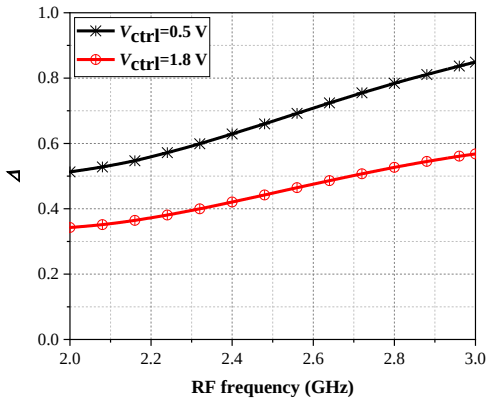
$$\Delta = |S_{11}S_{22} - S_{12}S_{21}| < 1 \quad (25)$$

Table 2: Simulation results of the proposed mixer with temperature and process corners.

Parameters	CG(dB)	NF(dB)	IIP3(dBm)	S_{11}(dB)
FF@-40	23.77	4.15	-9.2	-16.48
TT@27	23.93	3.74	-9	-18.68
SS@85	16.95	5.025	-7.6	-20.35



(a)



(b)

Fig. 16: Simulated stability factor of the proposed VG-LM.

As can be seen, the proposed VG-LM satisfies the conditions for unconditional stability over the frequency band of interest at V_{ctrl}=0.5 V and V_{ctrl}=1.8 V. Also, Table 2 presents the simulation results in different corners of the process and temperature changes with V_{ctrl}=0.5 V. As can be seen, in the worst case, the conversion gain drops to 16.95 dB because it strongly relies on the transconductance of M_{11}. The simulated performance of the proposed VG-LM is summarized in Table 3, and compared with the state-of-the-art results.

As can be seen, the proposed circuit benefits from high conversion gain, good input match, and low NF when it works in the maximum conversion gain. Moreover, it exhibits a wide conversion gain range and slight variations in the NF.

The following figure of merit (FoM) is used to have a fair comparison as follows:

$$FoM = 10\log\left(\frac{10^{\frac{CG_{max}(dB)}{20}} \times 10^{\frac{GR(dB)}{20}}}{10^{\frac{NF_{min}(dB)}{10}} \times P_{DC}(mW)}\right) \quad (26)$$

where P_{DC} and GR represent the power consumption and gain range, respectively.

As seen in Table 3, the proposed VG-LM has a high conversion gain compared to the recent structure as well as low DSB NF.

It results in an outstanding FoM proving the effectiveness of using the proposed LNTA in the RF stage of the Gilbert cell mixer. Although the IIP3 is not as high as [13], [25] and [27], it is comparable with [15], [16] and [29], which is acceptable for using in WLAN applications. The mixer reported in [13] exhibits a wide gain range, but it has low conversion gain. Moreover, some RF front-end circuits [26], [28] and [30] are evaluated to present a complete comparison with recent studies.

However, these circuits do not realize the variable gain and FoM is not reported for them.

Barzgari et al. [26] proposed a quadrature and differential RF front-end receiver for low power applications. By combining balun, LNA, mixer, and oscillator in a single stage, the proposed circuit features high integration level and low power consumption. Vitee et al. [28] presented an inductively source degenerated balun-LNA mixer.

They achieve good linearity by two linearization techniques but the NF and chip are are high in comparison with other designs. Bae et al. [30] proposed a new reconfigurable front-end circuit by using a reconfigurable parallel mixing subharmonic (SHM)-based time-interleaved RF channelizer. The circuit achieves high conversion gain but the linearity is the lowest among the RF front-end designs.

Table 3: Performance summary of the proposed VG-LM and comparison with previous works

Ref.	[13]	[16]	[17]	[18]	[19]	[20]	[25]	[26]	[27]	[28]	[29]	[30]	**This work**
CMOS Tech. (µm)	0.18	0.18	0.18	0.18	0.18	0.13	0.13	0.18	0.18	0.13	0.13	0.065	**0.18**
Topology	Mixer +VGA	Mixer +VGA	LNA +Mixer	LNA +Mixer	LNTA +Mixer	RX	Mixer +VGA	RX	Reconfig. Mixer	Balun+ LNA+ Mixer	Mixer +VGA	Reconfig. LNA+Mixer	**LNTA +Mixer +VGA**
Maximum CG[a] (dB)	8	24.2	22.28	20	17.87	45	17	57	19.67	22	23.7	46.7	**23.9**
Measurement Method	Sim.	Meas.	Sim.	Sim.	Sim.	Meas.	Meas.	Meas.	Meas.	Meas.	Meas.	Meas.	**Sim.**
LO Power (dBm)	0	-8	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0	N/A	0	**0**
RF Frequency (GHz)	1.4-1.6	1.8	2.4	2-10	2.4	5.8	1-12	2.2-2.8	0.7–2.3	2.2-3	1-6	0.3-0.8	**2.2-2.6**
IF Frequency (MHz)	N/A	10	595	10	N/A	5	110	2	10	100	140	3-10	**10**
Minimum DSB NF[a] (dB)	N/A	14	1[d]	6.8	5.9	<15	11	10.5	8.03	7.2	3.8	5.2	**3.74**
Chip area (mm^2)	0.02[b]	0.93[c]	N/A	0.52[b]	0.1	0.75	0.105[b]	0.75	0.71[c]	1.16	0.43[b]	3.6[c]	**0.57[c]**
Supply voltage (V)	1.8	0.8	1.8	1.8	1.8	1.5	1.2	0.8	1.8	1.2	1.5	1.2	**1.8**
P_{DC}[a] (mW)	7.56	2	37.4	7.2	9	33	5.9	0.34	23.76	3.15	N/A	24.7	**22.46**
IIP3[a] (dBm)	7	-11	N/A	-1	11.83	-44	8.6	-15.5	8.5	16	-4	-22.3	**-9**
S_{11}[a] (dB)	N/A	N/A	N/A	-7	-9	-23.7	N/A	<-10	N/A	<-10	-15	N/A	**<-10**
GR (dB)	45	9.7	N/A	3	N/A	N/A	15.8	N/A	18	N/A	13	6.4	**27.8**
FoM	N/A	-0.06	N/A	-3.8	N/A	N/A	-2.3	N/A	-2.93	N/A	N/A	7.42	**8.59**

a. High gain
b. Core
c. Core + Pads
d. only LNA

Conclusion

In this paper, a merged LNA-mixer with 27.8 dB variable gain range is presented. Using a low noise transconductance amplifier (LNTA) in the RF stage of the active mixer, a wide conversion gain range is realized without significant degradation in the noise figure. The proposed variable conversion gain LNA-mixer (VG-LM) is designed and simulated in RF-TSMC 0.18 µm CMOS technology. The post-layout simulated results exhibit an input matching (S_{11}) less than -10 dB, the maximum conversion gain of 23.9 dB, and the minimum DSB noise figure of 3.74 dB at the input frequency of 2.4 GHz. The simulated results demonstrate that the proposed VG-LM could be suitable for the low noise and wide tunable gain range RF front-end receivers in WLAN applications.

Author Contributions

M. A. Mallaki and A. Bijari developed the theoretical idea and performed the analytic calculations. M. A. Mallaki carried out the simulations. All authors discussed the results and contributed to the final manuscript. A. Bijari supervised the project.

Acknowledgment

We thank our colleagues from the university of Mashhad (FUM) who provided insight and expertise that greatly assisted the research. We thank M. Forouzanfar assistance for comments that greatly improved the manuscript.

Conflict of Interest

The author declares that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Abbreviations

CMOS	Complementary Metal Oxide Semiconductor
LNA	Low Noise AmplifierMultispectral
VG-LM	Variable Conversion Gain LNA-Mixer
WLAN	Wireless Local Area Network
LNTA	Low Noise Tansconductance Amplifier
DSB-NF	Double-Sideband Noise Figure
VGA	Variable Gain Amplifier
DCR	Direct Conversion Receiver
BPF	Bandpass Filter
LPF	Lowpass Filter
CMFB	Common-Mode Feedback
ADC	Analog to Digital Cconverter
NF	Noise Figure
AGC	Automatic Gain Control l
VCG	Variable-Conversion Gain
SFDR	Spurios Free Dynamic Range
DSRC	Dedicated Short Range Communications
MGTR	Multi-Gated Transistor
RF	Radio Frequency
IF	Intermediate Frequency
LO	Local Oscillator
CG	Conversion Gain
GR	Gain Range

References

[1] R. Mahmou, K. Faitah, “High linearity, low power RF mixer design in 65nm CMOS technology,” AEU Int. J. Electron. Commun., 68(9): 883-888, 2014.

[2] J. Borremans, P. Wambacq, C. Soens, Y. Rolain, M. Kuijk, “Low-area active-feedback low-noise amplifier design in scaled digital CMOS,” IEEE J. Solid-State Circuits, 43(11): 2422-2433, 2008.

[3] W. Chong, H. Ramiah, G. Tan, N. Vitee, J. Kanesan, “Design of ultra-low voltage integrated CMOS based LNA and mixer for ZigBee application,” AEU Int. J. Electron. Commun., 68(2): 138-142, 2014.

[4] J. Chen, W. Wang, “A K-band low-noise and high-gain down-conversion active mixer using 0.18-µm CMOS technology,” Wireless Pers. Commun., 104(1): 407-421, 2018.

[5] N. Seyedhosseinzadeh, A. Nabavi, “Design of an active CMOS subharmonic mixer with enhanced transconductance,” AEU Int. J. Electron. Commun., 73(1): 98-104, 2017.

[6] D. M. Pozar, Microwave and RF Design of Wireless Systems, John Wiley & Sons, 2000.

[7] M. Heping, H. Xu, B. Chen, Y. Shi, "An ISM 2.4 GHz low power low-IF RF receiver front-end," J. Semicond., 36(8), 2015.

[8] S. Amirabadizadeh, A. Bijari, H. Alizadeh, N. Mehrshad, "Performance improvement of a down-conversion active mixer using negative admittance," Circuits Syst. Signal Process., 40(1): 22-49, 2021.

[9] T. J. Rouphael, Wireless Receiver Architectures and Design: Antennas, RF, synthesizers, mixed-signal, and digital signal processing, Academic Press, 2014.

[10] A.D. Gungordu, N. Tarim, "Design of a constant-bandwidth variable-gain amplifier for LTE receivers," Analog Integr. Circuits Signal Process., 97: 27-38, 2018.

[11] H. D. Lee, K. A. Lee, S. Hong, "A wideband CMOS variable gain amplifier with an exponential gain control," IEEE Trans. Microwave Theory Tech., 55(6): 1363-1373, 2007.

[12] M. A. Martins, L. B. Oliveira, J. R. Fernandes, "Combined LNA and mixer circuits for 2.4 GHz ISM band," in Proc. 2009 IEEE International Symposium on Circuits and Systems: 425-428, 2009.

[13] C. W. Ryu, C. S. Cho, J. W. Lee, J. Kim, "A low power 45 dB dynamic-range variable gain mixer in 0.18µm CMOS," in Proc. 2009 IEEE MTT-S International Microwave Symposium Digest: 585-588, 2009.

[14] V. Kolios, G. Kalivas, "A 60 GHz down-conversion mixer with variable gain and bandwidth in 130 nm CMOS technology," in Proc. 2016 5th International Conference on Modern Circuits and Systems Technologies (MOCAST): 1-4, 2016.

[15] M. Wang, C. E. Saavedra, "Reconfigurable broadband mixer with variable conversion gain," in Proc. 2011 IEEE MTT-S International Microwave Symposium: 1-4, 2011.

[16] C. Wu, H. Chou, "A 2.4 GHz variable conversion gain mixer with body bias control techniques for low voltage low power applications," in Proc. 2009 Asia Pacific Microwave Conference: 1561-1564, 2009.

[17] C. Kalamani, "Design of differential LNA and double balanced mixer using 180 nm CMOS technology," Microprocess. Microsyst., 71: 102850, 2019.

[18] B. Hu, X. Yu, L. He, "A Gm-boosted and current peaking wideband merged LNA and mixer," in Proc. 2010 IEEE International Conference on UltraWideband (ICUWB): 1-4, 2010.

[19] S. C. Gladson, K. Alekhya, M. Bhaskar, "A fully CMOS RF down-converter with 81.88 dB SFDR for IEEE 802.15.4 based wireless systems," Microsyst. Technol., 28: 745-760, 2022.

[20] L. Cao et al., "A 5.8 GHz digitally controlled CMOS receiver with a wide dynamic range for Chinese ETC system," IEEE Trans. Circuits Syst. II Express Briefs, 65(6): 754-758, 2018.

[21] B. Guo, J. Gong, Y. Wang, "A Wideband differential linear low-noise transconductance amplifier with active-combiner feedback in complementary MGTR configurations," IEEE Trans. Circuits Syst. I Regul. Pap., 68(1): 224-237, 2021.

[22] B. Razavi, RF Microelectronics, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2011.

[23] A. Bijari, S. Zandian, "Linearity improvement in a CMOS down-conversion active mixer for WLAN applications," Analog Integr. Circuits Signal Process., 100: 483-493, 2019.

[24] C. H. Chen, M. J. Deen, "Channel noise modeling of deep submicron MOSFETs," IEEE Trans. Electron. Devices, 49(8): 1484–1487, 2002.

[25] J. Xu, C. E. Saavedra, G. Chen, "A 12 GHz-Bandwidth CMOS mixer with variable conversion Gain capability," IEEE Microwave Wireless Compon. Lett., 21(10): 565-567, 2011.

[26] M. Barzgari, A. Ghafari, M. Meghdadi, A. Medi, "A current re-use quadrature RF receiver front-end for low power applications: blixator circuit," IEEE J. Solid-State Circuits, 57(9): 2672-2684, 2022.

[27] X. Fan, J. Tao, K. Bao, Z. Wang, "A reconfigurable passive mixer for multimode multistandard receivers in 0.18 µm CMOS," J. Semicond., 37(8): 085001, 2016.

[28] N. Vitee, H. Ramiah, P. I. Mak, J. Yin, R. P. Martins, "A 3.15-mW +16.0-dBm IIP3 22-dB CG inductively source degenerated balun-LNA mixer with integrated transformer-based gate inductor and IM2 injection technique," in IEEE Trans. Very Large Scale Integr. VLSI Syst., 28(3): 700-713, 2020.

[29] F. Jiang, C. E. Saavedra, "Codesign of Mixer-VGA downconverter blocks," Can. J. Electr. Comput. Eng., 38(3): 199-203, 2015.

[30] S. Bae, D. Kim, D. Kim, I. Nam, D. Im, "A reconfigurable passive mixer-based sub-ghz receiver front-end for fast spectrum sensing functionality," IEEE Trans. Circuits Syst. I: Regul. Pap., 68(2): 892-903, 2021.

Biographies

Abolfazl Bijari was born in Birjand, Iran in 1982. He received B.S. degree in Telecommunication Engineering and M.S. and Ph.D. in Electronics Engineering from Ferdowsi University of Mashhad (FUM), Iran in 2005, 2007, and 2013, respectively. He is currently an Associate Professor in Department of Electronics Engineering, University of Birjand, Iran. His research interests include RFIC design, microwave filters and MEMS-based devices.

- Email: a.bijari@birjand.ac.ir
- ORCID: 0000-0002-0552-0721
- Web of Science Researcher ID: AAP-6805-2020
- Scopus Author ID: 55000092200
- Homepage: https://cv.birjand.ac.ir/bijari/en

Mohammad Amin Mallaki Mohammad Amin Mallaki was born in Birjand, Iran in 1995. He received B.S. and M.S. degree in Electronics Engineering from Birjand University, Iran in 2017, 2020, respectively. Currently, he is Ph.D. candidate in Department of Electronics Engineering, University of Birjand, Iran. His current research interests include RF circuit design for wireless communications.

- Email: amin_mallaki@birjand.ac.ir
- ORCID: 0009-0008-2523-0634
- Web of Science Researcher ID: N/A
- Scopus Author ID: N/A
- Homepage: N/A

How to cite this paper:
A. Bijari, M. A. Mallaki, "A merged LNA-Mixer with wide variable conversion gain and low noise figure for WLAN direct-conversion receivers," J. Electr. Comput. Eng. Innovations, 12(1): 175-186, 2024.

DOI: 10.22061/jecei.2023.10124.679

URL: https://jecei.sru.ac.ir/article_1987.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Review paper

A Comprehensive Review on Blockchain Scalability

A. Matani, A. Sahafi *, A. Broumandnia

Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.

Article Info

Article History:
Received 30 July 2023
Reviewed 25 September 2023
Revised 05 October 2023
Accepted 05 November 2023

Keywords:
Blockchain
Scalability
Consensus
Sharding
Throughput

*Corresponding Author's Email Address: *sahafi@iau.ac.ir*

Abstract

Background and Objectives: Blockchain technology as a distributed and tamper-proof data ledger is attracting more and more attention from various fields around the world. Due to the continuously growing of the blockchain in both transaction data and the number of nodes joining the network, scalability emerges as a challenging issue.

Methods: In this survey, the existing scalability solutions in the blockchain are discussed under five categories including on-chain scalability, off-chain scalability, scalable consensus mechanisms, DAG-based scalability, and horizontal scalability through sharding. Meanwhile, the novelties they have created on the fundamental layers of the blockchain architecture are investigated.

Results: As a result, the advantages and disadvantages of the discussed mechanisms are pointed out, and a comparison between them in terms of different scalability metrics such as throughput, latency, bandwidth, and storage usage is presented. Therefore, this study provides a comprehensive understanding of the various aspects of blockchain scalability and the available scalability solutions. Finally, the research directions and open issues in each category are argued to motivate further improvement efforts for blockchain scalability in the future.

Conclusion: Scalability allows blockchain system to sustain its performance as it grows up. Lack of scalability has a negative effect on the mass adoption of the blockchain in practical environments. This paper presents a profound analysis of the existing scalability solutions, the issues and challenges they address, and the ones that are not resolved yet. Consequently, it inspires novel ideas for more scalable and efficient blockchains in the future.

Introduction

Blockchain is a distributed ledger that eliminates the need for any third-parties and enables the participating nodes in a peer-to-peer network to agree on every single entry in the data ledger using a consensus mechanism. Thanks to interesting features such as decentralization, immutability, transparency, and trustlessness, blockchain has turned into the most breakthrough technology and has been used in developing the applications of diverse domains. For example, many works take advantage of blockchain technology in finance [1], [2], E-Healthcare [3], [4], Internet of Things [5]-[7], supply chain [8], [9], Electricity management [10], [11], insurance [12] and voting [13], [14].

Numerous works proceed research to address different challenges associated with blockchain systems e.g. security [15], [16], decentralization [17], [18], scalability [19]-[22], query processing [23]-[25], blockchain indexing [26], [27] and so on. Generally, in order to carry out a robust project, it is essential to make a trade-off among three key properties of blockchain including decentralization, security, and scalability. Vitalik Buterins, one of the co-founders of Ethereum [28], claims that the blockchain systems can only have two out of these three properties and refers to it as a scalability

trilemma. It is obvious that security is a vital feature for blockchain and not able to be sacrificed for any two other properties. On the other hand, decentralization is an intrinsic feature of blockchain systems. Therefore, scalability remains a challenging feature that should be handled without compromising security and decentralization. Scalability enables blockchain to manage the growing number of requests effectively and retains its performance over different aspects e.g. throughput, latency, storage, and bandwidth usage, as it expands. Due to the distributed and agreement-based nature of the blockchain, its throughput in terms of Transactions Per Second (TPS) is considerably low in comparison with traditional databases. In addition, to participate in the consensus-making process, the nodes need a huge storage space to maintain a copy of the data ledger and sufficient bandwidth to communicate with other peers during the consensus process. Consequently, the mentioned scalability issues can become bottlenecks, hindering the widespread adoption of blockchain technology. A number of solutions have been proposed in the literature to cope with these issues. This survey reviews some of the top-cited research works addressing scalability issues and groups them into 5 categories: (1) on-chain scalability, (2) off-chain scalability, (3) scalable consensus mechanism, (4) Directed Acrylic Graph (DAG)-based scalability, and (5) horizontal scalability through sharding.

On-chain scalability strategies [29]-[38] are aiming to improve scalability by modifying the core features and elements of the blockchain like block [31], [38] or transaction structure [29], [30]. On the other hand, off-chain strategies [39]-[46] are designed to leave transaction processing outside the blockchain to save storage space and mitigate blockchain workload. For instance, Lightening Network [39] and Raiden Network [40] have adopted this strategy to enhance scalability. Scalable consensus mechanisms [47]-[55] refer to the methods that lead to agreement on a greater set of transactions in a shorter time. The fourth category, namely DAG-based scalability, points to the solutions in which instead of traditional blockchain, an alternative data structure named DAG is used [56]-[62]. Generally speaking, in these methods, the data ledger is modeled as a directed acyclic graph with vertices representing users/accounts and edges representing transactions among them. Hence, the transactions can be processed independently resulting in a significant increase in the throughput of the data ledger. Finally, horizontal scalability through sharding implies solutions that periodically partition blockchain nodes into subsets called "shards" and allow parallel processing of the transactions in shards. Sharding is the most promising approach towards improving scalability, and sharding-based protocols [63]-[72] have achieved a high improvement in throughput and other scalability criteria. In the following, related works are investigated in more detail. The contributions of this paper are as follows:

- First, to provide a background of the blockchain components, a layered architecture of the blockchain along with key components within each layer is discussed.
- Then, existing scalability solutions are organized into a taxonomy and their ways of improving the blockchain scalability besides their advantages and disadvantages are debated.
- In addition, the solutions of each taxonomic category are compared in terms of their key characteristics and scalability improvements including throughput, latency, storage, and bandwidth/ communication overhead.
- Finally, the remaining issues and future research directions for each category of the scalability solutions are individually outlined.

The remainder of this paper is as follows. The next section gives the preliminaries of the blockchain. After that, the existing surveys in blockchain scalability are reviewed and compared with this work. Then, the research methodology followed by this paper is explained. In the following sections, a taxonomy of the blockchain scalability solutions is presented, existing works are surveyed in detail and future research directions and open issues are discussed. Finally, the last section concludes the paper.

Blockchain: Preliminaries

In this section, aiming to achieve a better comprehension of the subsequent explanations, a general architecture of the blockchain along with some fundamentals is described. According to the abstraction layer model suggested by the authors in [73], the blockchain architecture is comprised of five layers: (1) data layer, (2) network layer, (3) consensus layer, (4) execution layer, and (5) application layer. In the following, the functionalities of these layers and key components within each layer are explained.

A. Data layer

The data layer in the blockchain architecture is responsible for data management in blockchain systems. The main focus of this layer is on the data structure, transaction model, and cryptographic mechanisms such as digital signature, Merkle tree and hash function, that ensure the security and integrity of information stored on the blockchain.

I) Types of Data Ledgers

From the perspective of data structure, the data ledgers are divided into two main categories: blockchain (e.g. Bitcoin [74] and Ethereum [28]) and DAG (e.g. IOTA [56] and Nano [57]).

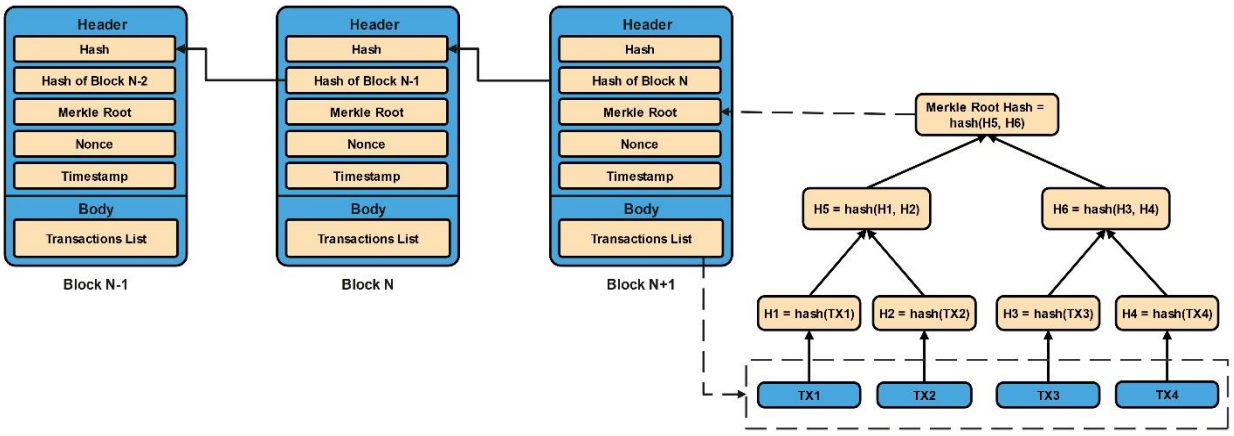


Fig. 1: Structure of blockchain [75].

Blockchain is a back-linked list of blocks chained together in an immutable and chronological order. As Fig. 1 [75] shows, to chain blocks together, each block is linked to the existing blockchain using the hash of the previous block. Each block consists of a set of verified transactions that are grouped together by a miner to be registered on the data ledger.

Opposite to the blockchains, in DAG-based ledgers, there is no chain of blocks and the data ledger seems like a graph. In other words, DAG is a network of individual transactions that are linked together and provide validation for each other. Practically, each new transaction must validate previous transactions and reference them to be registered on the network for validation. Therefore, the transactions that are directly or indirectly referenced by a given number of the transactions can be considered as committed.

Hence, there is no need for miners to mine blocks of transactions, resulting in fast confirmation times of the transactions and subsequently improving throughput and scalability. For example, in Fig. 2 [76], a weight is assigned to each transaction, and a transaction is considered as committed if the cumulative weights of the transactions which confirm it, be equal to or greater than 4 (as a threshold).

Other than blockchain and DAG, there exist other types of data ledgers that have been used in some data ledgers like Codra [77] and Radix [78].

II) Types of Transaction Models

The transaction is the main element for storing and exchanging information on the blockchain. Each transaction causes a blockchain transition from a valid state to another valid state.

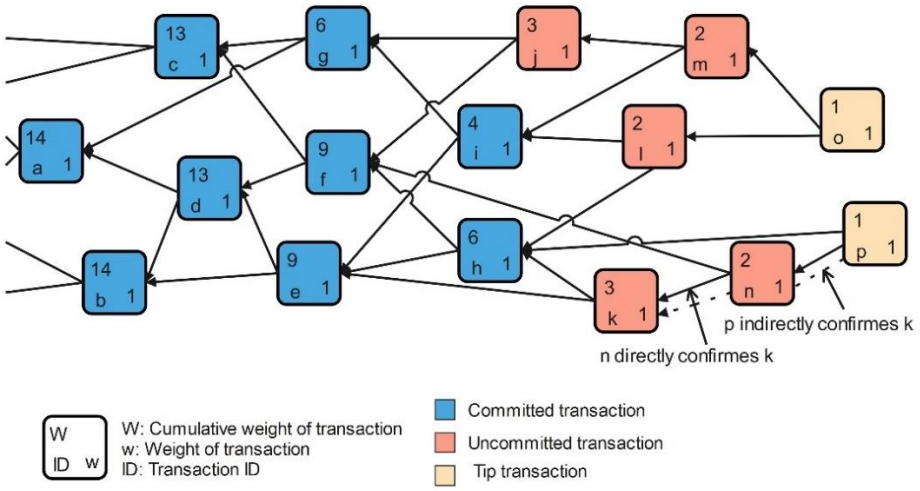


Fig. 2: An example of DAG structure named Tangle [76].

Two popular transaction models used in the blockchains are: Unspent Transactions Output (UTXO) and account-based. In the UTXO model, each transaction spends unspent outputs owned by the sender to create new outputs for a receiver as the new owner. In other words, assets owned by a user are scattered across the data ledger as unspent outputs of the transactions received by that user. The main advantage of the UTXO model is that it facilitates parallel processing of the transactions due to its atomicity and thus provides better infrastructure for scalability solutions. The bad point about the UTXO model is that it is only suitable for the applications in which each output is owned by one person. Moreover, it complicates the development of state-full smart contract-based applications because of its stateless nature.

In contrast, the account-based transaction model, analogous to the traditional banking model, maps each account into a balance. It has to be said that, the accounts' balances are stored in a global state trie that is constantly updated. Each transaction updates the global state as it deducts an amount from the balance of the sender and then adds that amount to the balance of the receiver. In comparison with the UTXO model, the account-based model is simpler and more efficient because transaction validation only needs to check whether the sender account has enough balance or not. In addition, it facilitates the development of smart account-based applications, specially state-full and multi-party ones. Nevertheless, scalability in account-based systems is more challenging than UTXO-based ones. Finally, it is worth mentioning that Bitcoin [74] uses the UTXO transaction model, whereas Ethereum [28] uses the account-based transaction model.

III) Cryptographic Components

In order to keep transactions data secure and immutable, blockchain uses cryptographic mechanisms, namely Merkel tree, hash function, digital signature, and Public Key Infrastructure (PKI) simultaneously. Merkle tree and hash function are used together to provide data integrity, whereas data signature is used to verify the authenticity of the transactions and ensure non-repudiation.

a) Merkle tree and Merkle Patricia Trie

In the blockchain, in order to chain the blocks in a tamper-resistant manner, each block header includes the hash of the previous block. Hence, any modifications in previously published blocks require changing all the subsequent blocks because they include the hash of the modified block.

In addition to the hash of the previous block that guarantees the integrity of the past transactions' history, each block contains a Merkle tree root that provides integrity for the current block's transactions (Reference to Fig. 1 [75]). Merkle tree is a binary tree in which each leaf node contains a hash of a transaction while each non-leaf node contains concatenated hashes of its children. Therefore, the Merkle tree root is a hash value obtained from the hash of all the transactions in the current block and any alternations in transactions will be detected by other nodes in the network because the Merkle root of the altered transactions will not match the one stored in the block header.

Bitcoin [74], in which transactions are the only state, uses the Merkle tree for the above-mentioned purposes. On the other hand, Ethereum [28], in which each node stores a global state consisting of a mapping between accounts and the account state, uses a Merkle Patricia Trie (MPT) that is an implementation of the Modified Merkle Patricia Trie [79]. MPT is a cryptographically authenticated key-value mapping that is used for storing and retrieving the accounts' state, as well as verifying data integrity. In MPT, leaf nodes store key-value states where the value is the account state and the key is its hash, whereas non-leaf nodes store the hash of the next node. Therefore, retrieving an account state needs to traverse MPT downward through the non-leaf nodes each of which stores the key of the next node, until reaching the leaf node storing the value corresponding to the searched key. The MPT allows checking data integrity by computing the Merkle root hash of the trie since if any key-value pair is modified, the Merkle root hash will not match for the entire list of the key-value pairs.

It must be pointed out that both the Merkle tree and MPT allow verifying the inclusion of a state (i.e. key-value state in MPT and transaction in Merkle Tree) without access to the entire blockchain using a method called Simplified Payment Verification (SPV).

After all, it is evident that the hash function is a fundamental component of blockchain technology. Most blockchains use the SHA-256 hashing algorithm, however, other hashing algorithms such as SHA-3 and Ripemd160 have been used by several blockchains.

b) Public Key Infrastructure and Digital Signature

In PKI technology, each user owns a pair of keys: a public key and a private key, which are used for authenticating users and protecting sensitive data. The public key is distributed on the network and is known to other nodes while the private key must be kept secret to never be known by any other nodes except its owner. PKI has algorithms that enable participating nodes in a network to encrypt, decrypt, sign and verify messages using their pairs of keys.

In PKI, if a message is encrypted with one key, it can only be decrypted with the second key. Therefore, if a message is encrypted with the public key of the receiver, it can only be decrypted with the private key of the receiver. In this case, the encrypted message is protected

from eavesdropping by malicious users since the receiver account is the only one that knows its private key and can decrypt the message. On the other side, if a message is encrypted with the private key of the sender, it can only be decrypted with the public key of the sender. In this case, the encrypted message is authenticated in terms of its source because the sender account is the only one who knows its private key and can encrypt that message. Hence, if a message is encrypted by the private key of the sender, the entire encrypted message serves as a digital signature since it ensures a receiver that the message has been encrypted by a claimed sender.

The digital signature is a primary usage of PKI technology in the blockchain. In fact, a digital signature is a mathematical function used to present the authenticity of the transactions and ensure non-repudiation and data integrity. Therefore, each transaction is signed by the private key of the sender to be authenticated by other participating nodes in the blockchain. Elliptic Curve Digital Signature Algorithm (ECDSA) is the most widely used data signature algorithm that has been used by Bitcoin [74] and numerous blockchain applications, although some blockchains use different digital signatures such as Edwards-curve Digital Signature Algorithm (EdDSA) [80], Borromean Ring Signature (BRS) [81], and One-Time ring Signature (OTS) [82].

In addition, a number of works [63], [66] use PKI combined with the Proof of Work (PoW) to establish identities for users securely and unpredictably.

B. Network Layer

Blockchain operates on a peer-to-peer (P2P) network that allows nodes to join the network and communicate with each other in a trustless way. The network layer is characterized by P2P network topology, peer discovery, identity management, block and transaction propagation, and takes care of privacy, anonymity, communication cost, security and attack resiliency.

I) Types of Nodes

There are two types of nodes in the network layer: lightweight nodes and full nodes. Lightweight nodes only store block headers and verify transactions by the SPV method, which allows users to verify the inclusion of a transaction in a block using a Merkle path and referencing to a trusted full node, whereas, full nodes store a complete and up-to-date copy of the blockchain and verify transactions autonomously without any external references. A full node is more reliable and safer than a lightweight node, however, it needs more storage space, bandwidth and computing power than a lightweight node.

II) Types of Blockchains

From the perspective of accessibility, blockchains are classified into two primary types: public and private. A public blockchain is open to the public so everyone can join the network. On the other hand, a private blockchain is closed to the public and each user requires to be authorized for joining the network.

Additionally, from the perspective of permission, blockchains are divided into two types: permission-less and permissioned. In a permission-less blockchain, each participating user can read, write or validate transactions without specific permission, whereas in a permissioned blockchain, authorized users need to obtain permission to read, write or validate transactions.

Finally, based on accessibility and permission, blockchains can be classified into four groups:

- Public permission-less (e.g., Bitcoin [74], Ethereum [28], Litecoin [83])
- Public permissioned (e.g., Ripple [84], EOS [85], Sovrin [86])
- Private permission-less (e.g., LTO [87], Holochain [88], Monet [89])
- Private permissioned (e.g., Hyperledger [90], Corda [77]).

C. Consensus Layer

The Consensus layer is a key aspect of the blockchain because in order to ensure consistency between the copies of the data ledger spread across the P2P network, the full nodes need to achieve a consensus on any updates to the data ledger. Essentially, the consensus process has an important role in many aspects of the system performance, such as scalability, integrity, and security. Consensus algorithms could be grouped into three following types: (1) proof-based, (2) vote-based, and (3) DAG-based.

In proof-based consensus algorithms, nodes compete to obtain the right to append the next block to the chain and the node that proves sufficient proof of qualification will win the competition. Proof of Work (PoW) and Proof of Stack (PoS) are the most popular proof-based consensus algorithms. For example, Fig. 3 [91] represents the flowchart of the PoW consensus process where the nodes need to prove their computational effort to add (mine) a new block to the network. To do so, a miner generates a random number (referred as to Nonce) and combines it with block data so that the hash value of the output data will be less than or equal to the current target of the network. The Proof-based consensus algorithms are appropriate for public blockchains since they provide high security in a trust-less system and also can easily scale in the number of users. Despite these advantages, this type of consensus algorithm reveals low transaction output and also the majority of them (e.g. PoW) are computation intensive and prone to the 51% attack occurring when a single node or a group of the nodes obtains control of more than 50% of the blockchain's

mining power.

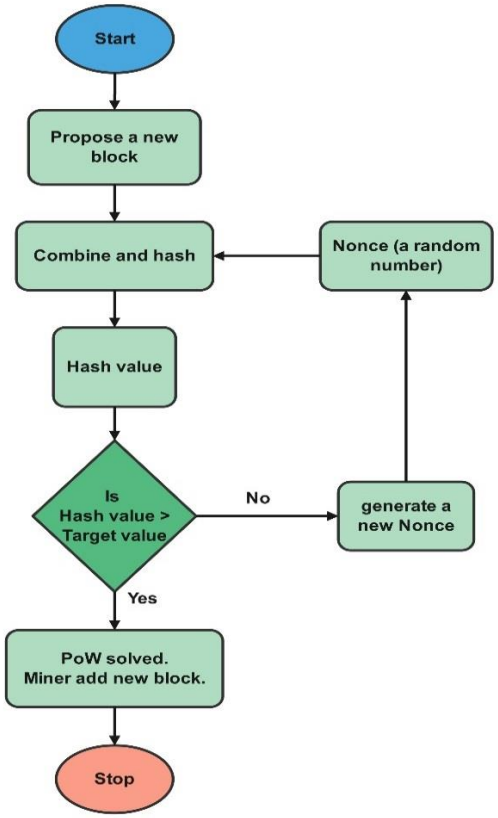


Fig. 3: Proof of Work (PoW) flowchart [91].

Oppositely, in the vote-based consensus algorithms, a leader is first elected to propose the next block. Then, the elected leader announces the next block to the other nodes having voting right. Afterward, each node participating in the voting process validates the proposed block and multicast necessary messages to the other nodes. Finally, if a given number of the nodes agree on a new block, it can be appended to the chain. Byzantine Fault Tolerance (BFT)-based algorithms such as Practical Byzantine Fault Tolerance (PBFT) [49] and Raft [92] are some popular examples of this type of consensus algorithm. The vote-based consensus algorithms have better transaction output and lower latency than the proof-based ones, although they are communication intensive and difficult to scale especially in large-scale environments. Hence, the vote-based consensus algorithms only work well on private and permissioned blockchains.

On the other hand, DAG-based consensus algorithms are used in the data ledgers that adopt DAG as their data structure, where a new transaction requires to validate the previous transactions in order to be processed by other transactions. In other words, in the DAG-based consensus algorithms, transactions provide validation for each other and can be processed in parallel, leading to fast transaction confirmation times. Fig. 2 [76] presented in Section "*Types of data ledgers*" depicts an illustration of a DAG-based data ledger where transactions are validated by each other. The DAG-based consensus algorithms have a comparative advantage in performance, scalability and simplicity, although their security can be compromised by malicious users who validate their transactions.

D. Execution Layer

The execution layer offers a runtime environment enabling nodes to participate in the network and interact with each other. A runtime environment is composed of Virtual Machines (VMs), compilers and containers that are installed on the computers and allow them to operate as a blockchain node. VMs contain APIs and services that enable nodes to execute and validate transactions, organize them into blocks and then share blocks with other peers.

Ethereum blockchain has developed its own virtual machine called Ethereum Virtual Machine (EVM). Ethereum nodes run EVM to execute smart contract code. A smart contract is a computer program that is executed automatically by nodes under predefined conditions. The smart contract helps transactions to be executed in a secure, transparent and conflict-free way. Smart contracts are written in a high-level language named Solidity [93]. Therefore, in order to run on an executing machine, smart contracts first need to be compiled into bytecode by the Solidity compiler, then these bytecodes are executed by EVM and deployed on the blockchain.

E. Application Layer

The application layer provides an interface for blockchain users to easily interact with the network, see results, share information and so on. In other words, the application layer is the first layer used by users to communicate over the blockchain network. Therefore, the usability and efficiency of blockchain applications greatly depend on the flexibility, speed and agility of this layer. Cryptocurrency providing a gateway for exchanging digital currency is the most popular example of the application layer. other examples are Decentralized Applications (DApps) developed in different domains and industries.

Existing Surveys

This section summarizes the existing surveys on blockchain scalability and outlines their contributions.

Hafid et al. [94] surveyed blockchain scalability under two categories: first-layer and second-layer. To enhance scalability, first-layer solutions modify the core features of the blockchain, whereas second-layer solutions are implemented outside of the blockchain and built on top of it.

In [94], sharding-based solutions along with other solutions including DAG-based and bigger block solutions are placed in the first-layer solutions, although the focus

is more on the sharding-based solutions. Hafid et al. [94] presented a taxonomy of sharding-based solutions based on committee formation and intra-committee consensus. However, Hafid et al. [94] reviewed a comprehensive range of the scalability solutions, they did not discuss the solutions in enough detail (specially sharding-based solutions).

Another disadvantage is that the discussed future works do not cover all the solution types and are only regarding sharding-based solutions.

Zhou et al. [95] presented a review of the blockchain scalability solutions and classified them into 3 layers: (1) layer 1 solutions, which are interrelated to the block data, consensus strategies, sharding, and DAG-based data ledgers, (2) layer 2 solutions that are associated with non on-chain techniques and include payment channel, side-chain, cross-chain, and off-chain computation mechanisms, and (3) layer 0 solutions whose main concern is to optimize data propagation in the blockchain. Despite the careful subdivision of the solutions in these layers, some key related works are not investigated in detail and their main contributions are not well defined. The main weakness of the survey [95] is that it lacks a comprehensive comparison of the discussed works and only a few methods are compared in terms of the throughput and confirmation time.

Differently from the aforementioned surveys [94], [95] the survey conducted in this paper provides a more detailed taxonomy of the blockchain scalability solutions including 5 categories: (1) on-chain scalability, (2) Off-chain scalability, (3) scalable consensus mechanisms, (4) DAG-based scalability, and (5) Horizontal scalability through sharding.

Exploiting such accurate taxonomy helps to discriminate and compare the key features of the various solutions precisely. Therefore, in this paper, a comprehensive comparison is presented separately for each category. In addition, the future work for each category is highlighted individually. Another advantage of this survey is that it compares the discussed solutions in terms of various scalability measurements including throughput, latency, storage usage, and communication overhead.

Nasir et al. [96], presented a systematic survey in which they define two dimensions for blockchain scalability: horizontal scalability and vertical scalability. Horizontal scalability refers to scaling blockchain by adding more nodes and clients, whereas vertical scalability refers to boosting the capabilities of the participating nodes such as processing power, storage capacity, memory, and efficient strategy. Vertical scalability is further broken down into several sub-dimensions including throughput, latency, block generation rate, and storage (chain size and block size). Nasir et al. [96] categorized scalability solutions into 5 groups including: (1) on-chain solutions, (2) off-chain solutions, (3) hardware-assisted approaches, (4) parallel mining/ processing, and (5) Redesigning blockchain, although they did not go deep into the solutions of each category. In addition, scalable consensus mechanisms have not been investigated in the survey [96].

Yu et al. [97] provided a survey focusing on the sharding solutions. They presented a comprehensive comparison of the key features of the sharding-based solutions and also conducted a systematic analysis of the scalability metrics such as throughput, latency, storage and communication complexity, although the debated solutions are limited to only a small number of sharding-based solutions.

Wang et al. [98] provided an overview of state-of-the-art DAG-based blockchains and also abstracted a general model to describe them in a theoretical and mathematical form and then identified 6 types of DAG-based blockchain systems.

They evaluated and compared the studied systems from the perspectives of their structure, consensus mechanism, security, and performance (in terms of scalability, throughput, and latency).

Oyinloye et al. [99] presented a comprehensive overview of the alternative consensus protocols which have been proposed in recent years, even the lesser-known ones. They evaluated the alternative consensus mechanisms in terms of throughput, scalability, security, energy consumption and block/ transaction finality (including absolute/ immediate finality and probabilistic finality).

Therefore, the main advantage of this survey over tree above-mentioned surveys [97]-[99] is that it covers a comprehensive variety of scalability solutions, instead of focusing only on the sharding-based solutions or DAG-based solutions and alternative consensus protocols. Table 1 presents a summary comparison between this work and the described surveys.

Research Methodology

This survey is accomplished based on four Research Questions (RQ) and is aiming to answer these questions at the different steps of the study. The questions are as follows:

- RQ1: What are the scalability bottlenecks in the blockchain systems?
- RQ2: What metrics are used to measure blockchain scalability?
- RQ3: Which blockchain elements can be manipulated to improve scalability?
- RQ4: What are the open issues and future prospects for the blockchain scalability?

The research methodology of this survey consists of 6 steps that are described below:

Table 1: Comparison Between this work and existing surveys

Reference	year	Publisher	Covered years	Covered Scalability solutions					Evaluation metrics
				On-Chain	Off-Chain	Consensus mechanism	DAG	Shrading	
This Work	--	--	2014-2021	✓	✓	✓	✓	✓	Throughput, Latency, Storage, Bandwidth
Hafid et al. [94]	2020	IEEE	2014-2020	✓	✓	✓	✓	✓	Throughput, Latency
Zhou et al. [95]	2020	IEEE	2014-2019	✓	✓	✓	✓	✓	Throughput, Latency
Nasir et al. [96]	2021	Elsevier	2015-2020	✓	✓	×	✓	✓	Throughput, Latency, Block generation rate, Storage
Yu et al. [97]	2020	IEEE	2016-2019	×	×	×	×	✓	Throughput, Latency, Storage and Communication complexity
Wang et al. [98]	2020	arXiv preprint	--	×	×	×	✓	×	Scalability, Throughput, Latency
Oyinloye et al. [99]	2021	MDPI	2018-2020	×	×	✓	×	×	Throughput, Scalability, Security, Energy consumption, Finality

F. Keywords Generation

To generate the keywords for searching the relevant research works, first, some possible answers were provided for RQ1 and RQ2.

Then, two sets of keywords were extracted from the answers of the RQ1 and RQ2, respectively named K1 and K2.

As can be seen in Table 2, K1 and K2 keyword sets were combined with K, a keyword set including primary keywords such as blockchain, scalability, scalable and scaling, to generate the expressions for searching among electronic databases (considering the synonyms words).

Table 2: Process of generating the keywords and searching expressions

	Set	Keywords
Main Keywords for blockchain scalability	K	blockchain, scalability, scalable, scaling
RQ1: What are the scalability bottlenecks in the blockchain systems?	K1	network size, blockchain size, high communication overhead, storage, block size, consensus (inefficient consensus strategies)
RQ2: what metrics are used to measure blockchain scalability?	K2	throughput, latency, storage usage, bandwidth
Search expressions = ((k1 or k2) and k) where {k ∈ K, k1 ∈ K1, k2 ∈ K2}		

G. Searching of Research Works

In this phase, using the generated search expressions, the research works were searched in electronic databases such as IEEE, Springer, ACM, Elsevier, Google Scholar, Taylor & Francis and so on. At last, 137 research works containing the mentioned keywords were found.

H. Refinement of Research Works

To select more relevant and valuable research works, among the 137 discovered researches, the ones having the below conditions have been excluded from the study:

- The papers not written in English.
- The papers published before the year 2014 (except the highly cited ones).
- Short papers with less than 8 pages (except the highly cited ones).
- The low-citation papers that have been published before the year 2019.
- The review papers.

The output of this phase is 33 research papers.

I. Cross Checking the Selected Research Papers

In this phase, the references of the selected papers in the previous phase were checked out, to ensure not missing the important and valuable researches.

This checkout resulted in finding 8 other papers. Therefore, during this study, totally 41 research works in the field of blockchain scalability have been studied in

detail.

J. Classification and Review

Finally, the selected papers were studied and categorized based on the strategy that they have applied to enhance blockchain scalability.

K. Identification of Future Research Directions

Meanwhile evaluating the research works, the unresolved issues and also some promising directions were identified and have been recommended at the end of this study for future works.

Fig. 4 illustrates a summary of the steps followed by the methodology of this survey in sequential order. Fig. 5 and Fig. 6 also respectively show distribution of the reviewed research works by publisher and publication year.

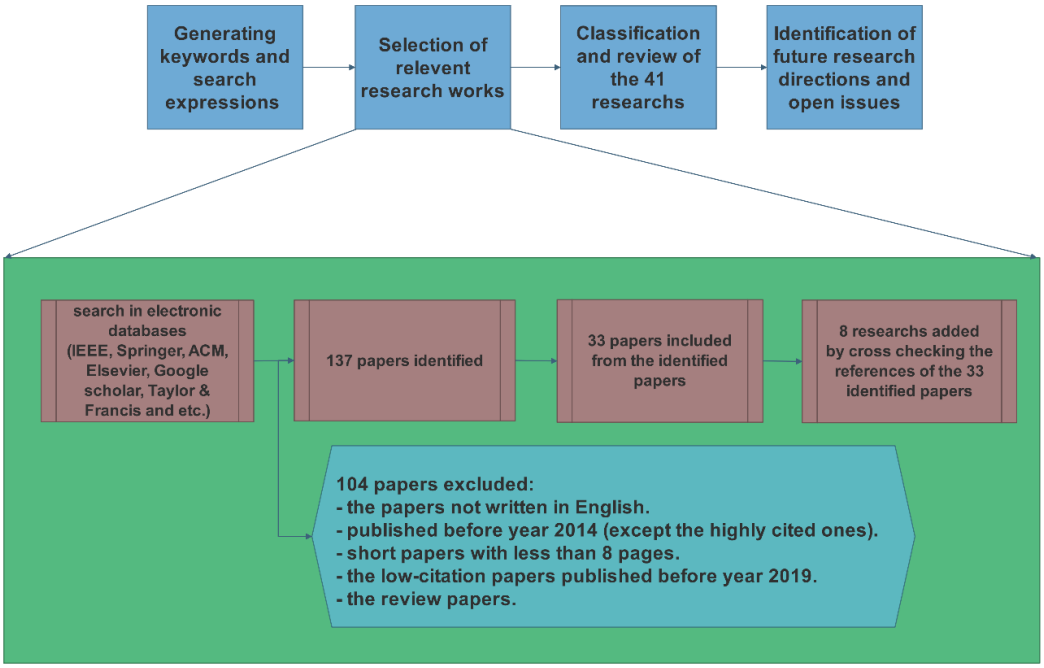


Fig. 4: Summary of research methodology.

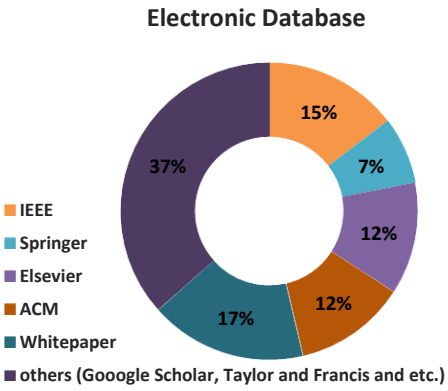


Fig. 5: Distribution of research works based on electronic database.

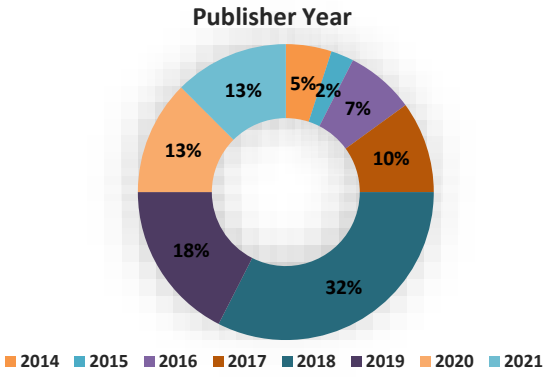


Fig. 6: Distribution of research works based on publish year.

Scalability Issue

With the continuous growth of blockchain systems, scalability is emerging as a challenging issue and the biggest barrier to the widespread adoption of the blockchain.

Despite the increasingly growing scale of the system, a scalable blockchain not only retains its functionality and performance but also takes advantage of the larger scale system to improve its performance. Ever-increasing transaction data and number of the participating users in blockchain systems lead to scalability issues such as low

throughput (in terms of transactions per second), latency and also the greater need for storage and bandwidth. Since blockchain is a consensus-based and distributed data ledger where all the full nodes keep a copy of the data ledger and validate all the transactions, the more users join the network, the more time is needed to reach a consensus on the transactions. Therefore, latency increases and the overall throughput of the system decreases. Moreover, with the growing size of the blockchain, the full nodes require more storage space to replicate the data ledger and more bandwidth to download the whole data ledger to bootstrap at initialization time. In this survey, scalability solutions are grouped into five categories as follows:

- On-chain scalability
- Off-chain scalability
- Scalable consensus mechanisms
- Directed Acrylic Graph (DAG)-based scalability
- Horizontal scalability through sharding

Fig. 7 demonstrates the taxonomic categories and the existing solutions in each category that have been discussed in this paper. Each category of scalability solutions can make changes on the different layers in the blockchain architecture. In the following sections, a detailed survey of the existing scalability solutions is presented. Meanwhile, these solutions are analyzed from the perspectives of scalability metrics including throughput, latency, storage, and bandwidth.

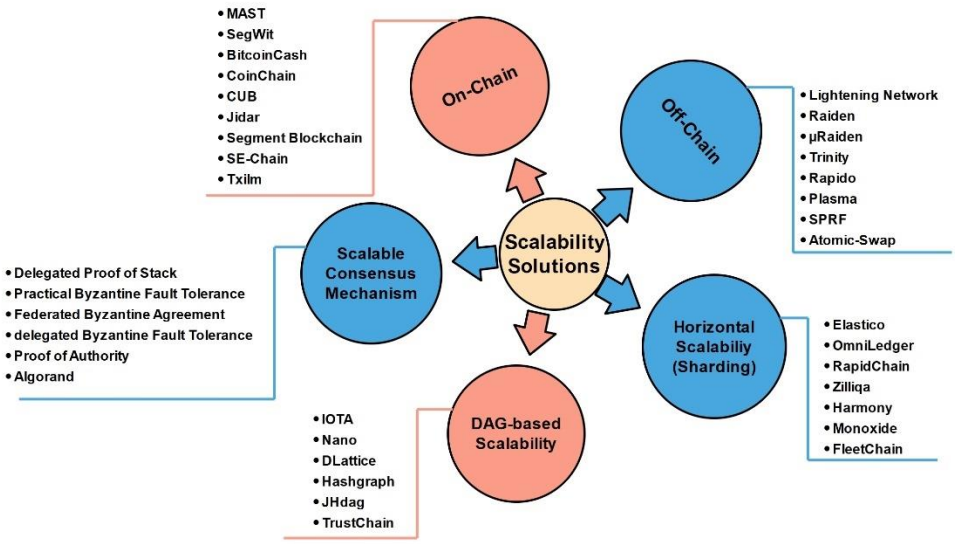


Fig. 7: Taxonomy of scalability solutions.

L. On-Chain Scalability Solutions

On-chain scalability solutions refer to the solutions that modify some key elements of the data layer in the blockchain architecture. According to the Data layer section of this paper, key elements of the data layer are block, transaction, Merkle tree, digital signature, and hash function.

For example, some works [31], [100] use a bigger block method in which the block's size is increased to a larger size. The bigger block method provides a higher throughput because bigger blocks can contain more transactions, thus whenever a block is added to the blockchain more transactions are confirmed, although the bigger blocks lead to a higher block propagation delay.

In addition, block compression is another method that is used by some works [101], [102] to save both the space of the blockchain and the bandwidth of the network. In this section, some of the important on-chain scalability solutions are introduced.

I) Merkelized Abstract Syntax Tree (MAST)

Merkelized Abstract Syntax Tree (MAST) [29], as an addition to Bitcoin, has been proposed to improve the scalability of the blockchain from the aspect of the capacity of the Bitcoin scripting. To do so, it combines the Merkle Tree and Abstract Syntax Tree (AST) concepts to represent the script parts of the Bitcoin transactions, both compactly and securely. Indeed, the transaction outputs in Bitcoin include a locking script also known as "encumbrance" that specifies the conditions under which the recipient can spend that output. Furthermore, the AST is a tree that represents the abstract syntax of the source code of a computer program as a hierarchical tree structure. Therefore, by employing AST, MAST is able to store the more complex locking scripts in Merkle tree format and remove unused parts of a script from the transaction. Thus, MAST combining the Merkle tree and AST, provides both data integrity and transaction compression. To sum up, MAST causes smaller

transactions and more privacy, and allows larger smart contracts, however, it increases the complexity of the contracts agreed to the Bitcoin.

II) Segregated Witness

Segregated Witness (Segwit) [30] is a modification to Bitcoin [74], whose solution for scalability is to free up space on the blocks in order for more transactions can be included in a block, whereby more transactions would be carried out and transaction confirmation speeds up.

Bitcoin transactions consist of inputs, outputs and also witness data including signature and script for transaction validation. To free up space on the block, Segwit removes witness data from the Bitcoin transactions and stores it onto a separate block instead of maintaining it on the blockchain. Since witness data consumes nearly 70% of the block size, by removing it, Segwit can process 1.7 to 4 times more transactions than Bitcoin, resulting in reduced transaction fees [103].

Segwit also aims to prevent transaction malleability in Bitcoin. Malleability refers to the problem that allows attackers to change the transaction ID of an existing transaction by modifying its digital signature, while its data is the same as the original one, and then rebroadcast it onto the network making other nodes think the original one has not been confirmed [104]. In addition to solving scalability and malleability issues, Segwit paves the way for developing off-chain solutions (e.g. Lightening Network [39] which will be discussed in the following sections) and now is utilized by Litecoin [83]. Having said all these advantages, the main disadvantage of Segwit is that, as a soft fork, it leads to a fungibility problem because there is no need for all the nodes in the network to upgrade the older version of Bitcoin. Additionally, Segwit extremely increases resource usage such as capacity and bandwidth because it needs to process more transactions at the same time. Furthermore, the implementation of Segwit is challenging since it increases code complexity.

III) BitcoinCash

Unlike Segwit soft work, BitcoinCash [31], [32] is a hard fork from Bitcoin that splits the Bitcoin network into two new blockchains. To improve scalability, Segwit tries to reduce transaction size while BitcoinCash tries to increase the block size. In fact, BitcoinCash changes the native Bitcoin codebase to increase the block size limit from 1MB to 8MB resulting in a higher throughput that is averagely 116 transactions per second. Consequently, BitcoinCash enables faster transaction processing than the Bitcoin network, while at the same time, it compromises decentralization because fewer nodes can process or propagate the larger blocks. In other words, it requires more processing capacity and bandwidth.

IV) CoinChain

CoinChain [33] is a scalable and prunable blockchain while keeping privacy and works as a sidechain for Bitcoin. Coinchain is scalable from the perspective of storage and blockchain size, whereas its transaction throughput is the same as Bitcoin. Indeed, privacy concerns to provide anonymous transactions and preserve confidentiality and anonymity of the sender, receiver and amount of the transactions, make the existing cryptocurrencies more complicated and restrain blockchain pruning. On the contrary, CoinChain is a straightforward and simple protocol that operates like physical cash transfer systems where banknotes with unique serial numbers, corresponding to distinct denominations, are transacted between users. Therefore, each coin is identified by a unique CoinID and coin ownership is transferred through transactions. Consequently, blockchain can be pruned by just keeping the last owner of the coins. Nevertheless, there are some shortcomings regarding CoinChain. First, fractional amount payment is not allowed. Second, the users are required to mix or spend out all the coins they pegged in the first place to ensure privacy. Third, in CoinChain, auditing for different purposes such as tracking money laundering or tax evasion, is feasible only via full disclosure of the transaction information.

V) Storage efficient solutions

High storage usage is a challenging issue that restricts many devices to participate in the blockchain because of storage space limitations. In this section, some works focusing on storage optimization in the blockchain will be discussed.

CUB [34], to reduce storage usage, splits the entire network into smaller units namely "consensus units" where the nodes cooperatively store one copy of the blockchain instead of each node keeping its own distinct copy. Therefore, it helps to save storage space for blockchain network peers. In addition, CUB provides solutions to optimize the block assignment and minimize the query cost. The main drawback of CUB is that it relies on a strong trust assumption which is hard to satisfy in practice.

Jidar [35] is a data reduction strategy without trust assumption for Bitcoin in which each node only has to store relevant transactions that it cares about, besides the branches of the Merkle tree from the whole block that is needed for the validation of new transactions. Jidar is able to reduce the storage cost of each node by about 1.03% compared with the native Bitcoin system. The bad thing about Jidar is that it does not support general-purpose smart contracts.

In addition, if some nodes require to have a whole block, they first need to query the pieces of the block data from different nodes and then cohere all pieces into a block, however, this functionality requires an incentive mechanism to be added.

Segment blockchain [36] is a data-reduced storage

approach for blockchain. The main idea of the Segment blockchain mechanism is that it partitions the blockchain into segments and then allows each node to store only one segment of the blockchain rather than the whole blockchain. It is proved that Segment blockchain reduces storage requirements significantly without compromising either the security or the decentralization of the blockchain. Furthermore, Segment blockchain facilitates blockchain sharding because it separates transaction verification from transaction storage. On the downside, it is only suitable for applications that do not need a large transaction output. SE-Chain [37] is also a scale-out blockchain framework that enhances storage scalability. In the data layer of the SE-Chain framework, each transaction is stored in the Adaptive Balanced Merkle tree (AB-M tree) and the full nodes store a part of the blockchain designated by the duplicate ratio regulation algorithm. In addition, to ensure the safety of the stored data on the full nodes, a node reliability verification method is presented. Another contribution of the SE-Chain is that it provides fast and efficient data retrieval using the AB-M tree.

VI) Block compression

Some works in the literature have used block compression to save the network bandwidth, an important factor that impacts blockchain scalability.

One such solution is Txilm [38] in which each block includes a list of compact presentation of the transactions rather than the original transactions. To produce a compact of a transaction, the transaction is hashed twice, first using SHA256 which generates a hash of 256-bits so-called TXID, then using a hash function (e.g. CRC32, CRC4p or CRC64) which generates a k-bit small-sized hash value so-called TXID-HASH. Therefore, the final output, i.e. TXID-HASH, is the compact presentation of the transaction that is included in a block along with the TXID-HASH of other candidate transactions and the block header which includes SHA256 Merkle root of all containing TXIDs. After that, the resulting compact block is propagated into the network by the user. Once receiving the compact block by full nodes, they should search into their memory pool to find a TXID matched with each TXID-HASH listed in the compact block. If one matched TXID is found, The TXID-HASH will be accepted. Otherwise, the full node requests the sender or other nodes for the missing TXID. Moreover, the hash collision that happens whenever multiple matches are found for a TXID-HASH, is resolved using Merkle root. As a final point, Txilm results in 80 times data reduction, thus saving the network bandwidth considerably and improving the blockchain throughput.

Comparison of On-Chain Scalability Solutions

To give a clear overview of the on-chain scalability solutions, a comparison of them is summarized in Table 3 where the mechanism used by each solution to improve scalability, and also the scalability metrics over which they achieve improvement are specified. Moreover, the advantages and disadvantages of each solution are neatly summarized.

Table 3: Comparison of on-chain scalability solutions

Solution	Mechanism	Scalability measurements				Advantages	Disadvantages
		Throughput	Latency	Storage	Bandwidth		
MAST [29]	Transaction compression	--	--	Low↓	--	- Smaller transactions - More privacy - Larger smart contracts	- Increases complexity of permitted contracts - Not complete privacy
SegWit [30]	Increased block capacity	High↑	--	--	High↑	- High transaction speed - Paves the way for developing off-chain solutions	- Results in fungibility problem - Increases processing capacity and bandwidth usage - Difficult to implement
BitcoinCash [31], [32]	Increased block size	High↑	--	--	High↑	- Increases throughput	- Compromises decentralization
CoinChain [33]	Blockchain Pruning	--	--	Low↓	--	- Simple and easily understandable - Secure with high privacy	- Fractional amount payment is not allowed - Users need to mix or spend out all the coins they initially pegged - Full transaction disclosure is needed for auditing
CUB [34]	Saving storage usage	--	--	Low↓	--	- Storage efficient	- Relies on a strong trust assumption
Jidar [35]	Data reduction	--	--	Low↓	--	- Storage efficient	- Does not support the general-purpose smart contracts
Segmant Blockchain [36]	Data-reduced storage	--	--	Low↓	--	- Storage efficient	- Suitable for the applications that do not need a large transaction output
SE-Chain [37]	Storage scalability	--	--	Low↓	--	- Efficient storage and data retrieval	--
Txilm [38]	Block compression	High↑	--	--	Low↓	- Saves bandwidth - Increases throughput	--

M. Off-Chain Scalability Solutions

The off-chain scalability refers to the solutions in which some portion of the transactions are offloaded from the blockchain to ease the burden of storing all the blockchain data in the main chain. Indeed, the off-chain transactions are executed outside the blockchain and only the final states are to be applied in the main chain. Consequently, they mitigate the issues arising from the ever-growing of the blockchain data thus improving the scalability and overall performance of the blockchain. Additionally, the off-chain transactions lead to lower fees and almost zero waiting time.

The off-chain mechanisms are usually in the form of payment channels [39]-[42] and sidechains [44]. A payment channel allows users to interact and transact with each other without using the expensive and slow blockchain and then broadcast the final closing transaction into the blockchain network to update their states. The payment channel is also called the state channel because it modifies and maintains the states of the main blockchain and then applies the last state to the main chain. On the other hand, a sidechain is an individual blockchain linked to its parent blockchain using a two-way peg [105] that allows users to interchange their assets between the sidechain and the parent chain at a prefixed rate.

In the following, some of the off-chain solutions that have attracted more attention are introduced.

I) Lightning network

The Lightening Network of Bitcoin [39] is one of the prominent examples of off-chain solution, which utilizes the payment channels to lighten the workloads of the main chain in Bitcoin. Therefore, every two users willing to transact with each other must first establish a channel between each other. For doing so, they first need to share a multi-signature address (wallet) and then they both deposit a certain amount of Bitcoin into that address. After that, they can do unlimited payment transactions between each other quickly and with minimal fees. After the transactions, the payment channel is closed and the final transaction is broadcasted to the Bitcoin blockchain to update the balance of the two users. The final transaction charges a fee from the payer user.

Despite all these advantages, Lightening Network has some drawbacks as follows: (1) it is less secure than the original Bitcoin, (2) it also supports only the micropayments for Bitcoin, and (3) it forces the users interacting with each other to be online at the same time and follow the same payment path.

II) Raiden Network and µRaiden

Raiden Network [40] is the Ethereum version of the Lightening Network, which allows Ethereum users to open a private channel namely "state channel" under which they can perform off-chain transactions and transfer tokens immediately and economically. The Raiden network is based on the same concepts as Lightening Network, however, opposite to the Lightening Network, it supports general-purpose transactions as the Ethereum supports general-purpose smart contracts.

µRaiden [41] is the first release of the Raiden network launched on the Ethereum mainnet. µRaiden is a micropayment solution for fast and free ERC20 token exchange. It is a many-to-one unidirectional payment channel framework that does not allow multi-hop transfers through payment channels, while Raiden is a many-to-many bidirectional solution that enables multi-hop transfers via bidirectional payment channels. Consequently, Raiden has a more complicated design than µRaiden.

III) Trinity

Trinity [42] is analogous to the Lightening Network and Raiden and provides an off-chain scaling solution using state channel technology. The difference is that it is built on the Neo blockchain [52] and aims to achieve real-time payment with low fees and provide protection for Neo assets. It increases the throughput of Neo blockchain considerably.

IV) Rapido

Rapido [43] is a scalable blockchain that provides multi-path payment channels, whereas before-mentioned off-chain solutions [39]-[42] offer a single-path payment. Single-path payments are vulnerable to leakage of sensitive information like payment value and also may result in overload issues since in the case that payment value goes beyond the deposit of every single path between two involved users, all the pre-established payment channels have to be closed and a new one needs to be established. Hence, Rapido has been presented to address these two issues by proposing Value Distributing Problem (VDP) program, whereby the payment value is divided into two or more sub-values and then sub-payments are settled using different payment paths. Furthermore, Rapido introduces Distributed Hashed Timelock Contracts (DHTLC) to ensure the security of these sub-payments. Rapido yields a success rate over 3 times higher than Lightening Network [39], although due to the existence of several intermediaries in multi-path payments, the willingness of individual donation is diminished [106].

V) Plasma

Plasma [44] provides an off-chain scaling solution for the Ethereum network through a sidechain mechanism. It employs a hierarchical tree-like structure of the chains called "child chains", that stem from the Ethereum main chain as its root (as shown in Fig. 8 [44]). Each child chain is a smaller version of the main chain and has its own subtrees of child chains. The child chains are smart contracts built on the main chain and have their own set

of rules and operate independently and in parallel to other chains. Therefore, child chains can be utilized for different purposes and interact with each other.

That is to say, the states maintained by the child chains are updated in the main blockchain periodically and verified by validating Merkle root. Plasma also allows the users to transfer their assets to the main blockchain by executing an "output transaction".

Taking advantage of these characteristics, Plasma can reduce the congestion of the Ethereum blockchain considerably and result in fast and low-cost transaction processing. Another advantage of Plasma is that it is consistent with other scalability solutions such as sharding and big blocks.

On the downside, to guarantee immutability in all child chains, many security considerations should be considered and addressed by the Plasma framework. Another challenge in Plasma is that in the case that, at the same time, all the users decide to leave the child chain and transfer their assets to the main chain, processing all the requests is impossible for the main chain.

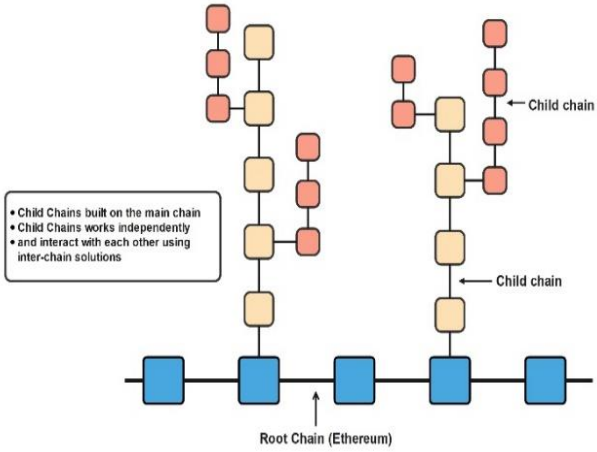


Fig. 8: Tree-like structure of Plasma blockchain [44].

VI) Smart Program Runner Framework (SPRF)

SPRF framework [45] is a sidechain implemented on the Stellar blockchain [50]. It is designed to move decentralized applications data off the blockchain and only stores the hash of the application state on the main chain. Therefore, it allows the state-tracking of smart programs on the blockchains. Indeed, it provides a platform including different applications, that enables decentralized compute-intensive software to be executed on the blockchain securely and efficiently. Another good thing about SPRF is that it can set up a sidechain for the existing blockchains without any need for revision.

It is important to highlight that there is another type of scalability solution named inter-chain which is similar to the sidechain. Inter-chain blockchains are used to connect different blockchains in a sidechain technology and solve interoperability problems between them. An inter-chain blockchain defines necessary protocols and standards enabling the blockchains to communicate among themselves. Atomic-swap [46] solutions are an example of inter-chain blockchain that provide an infrastructure for interacting between blockchains without the need for any centralized intermediaries, although they are applicable in bounded situations. For example, each blockchain must conform to extra programming features to be able to communicate with other blockchains.

Comparison of Off-Chain Scalability Solutions

In this section, a tabular comparison of the discussed off-chain scaling solutions is presented. As Table 4 demonstrates, these solutions have been investigated in terms of scalability improvements and grouped based on their off-chain mechanism. Moreover, their advantages and disadvantages have been outlined.

N. Scalable Consensus Mechanisms

There exist some scalable consensus mechanisms aiming to optimize the scalability and performance of blockchain systems. To achieve this goal, these mechanisms revolutionize the consensus layer to speed up the consensus-making process and subsequently increase the transaction throughput. In the Consensus layer section of this paper, a taxonomy of consensus algorithms and their underlying features has been presented. This taxonomy includes three types of consensus algorithms, namely Proof-based, Vote/ BFT-based and DAG-based.

DAG-based consensus algorithms used in the DAG-based data ledgers are potentially scalable, while two other types are not inherently scalable and scalability is a challenging issue in early consensus protocols developed based on them. Hence a variety of novel protocols have been proposed to renovate proof-based and vote/ BFT-based protocols.

For example, PoW is a primary proof-based protocol that has limitations regarding speed and scalability metrics such as throughput, latency, computational power and transaction capacity, although, it scales well in terms of network size and provides permission-less access to the network. Moreover, PoS was proposed as an alternative to PoW that mitigates some of its scalability issues such as high computational power, high latency and low throughput. On the other hand, BFT-based protocols have higher throughput than proof-based ones, but due to some internal drawbacks such as scalability and communication overhead, they are only suitable for a private network and need an identity management system.

Some proof-based consensus algorithms proposed in the literature are Proof of Stake (PoS), Proof of Authority (PoA) [53], [54], Proof of Elapsed Time (PoET) [107], Proof

of Capacity (PoC) [108], Proof of Importance (PoI) [109] and Proof of Burn (PoB) [110].

On the other side, Voting-based consensus algorithms are Delegated Proof of Stake (DPoS) [47], [48], Practical Byzantine Fault Tolerance (PBFT) [49], delegated Byzantine Fault Tolerance (dBFT) [51], [52], [111], Federated Byzantine Agreement (FBA) [50] and Algorand protocol [55]. In the following, some of the mentioned consensus algorithms are described briefly. Further details on consensus algorithms are available in [99].

I) Delegated Proof of Stack

Delegated Proof of Stake (DPoS) [47], [48] is an evolution of the PoS algorithm and allows the blockchain to reach consensus using a democratic manner. In DPoS, the stackers vote and elect delegates to validate the next block on their behalf. Delegates are also called validators or witnesses. During the voting process, stackers pool their stack into a staking pool and link them to a particular delegate. For each new block, between 20 and 100 delegates are chosen depending on the system. Chosen delegates add blocks to the chain in a Round-Robin manner. In fact, contrary to PoW and PoS which are competing systems, DPoS is a collaborative system where the delegates collaborate to make the blocks. The delegates that constantly miss their block or publish invalid blocks, will be voted out by stackers and replaced. The transaction fee for each validated block is shared between the stackers who elect the successful delegate. DPoS is partially centralized, however, it is more scalable than PoW and PoS. Furthermore, DPoS is more susceptible to 51% attack, because its consensus process depends on a small set of delegates.

EOS [85] is a blockchain technology, which utilizes DPoS to elect and schedule the block validators. Then, elected validators use an Asynchronous Byzantine Fault Tolerant (ABFT) consensus mechanism to validate and confirm the block proposed by the active validator and reach a consensus on it. EOS is aiming to enhance scalability and eliminate transaction fees. It also facilities the DApps development process. In addition to EOS, BitShares [48], Steemite [112], Ark [113], Cardano [114] and Lisk [115] are some of the well-known projects employing the DPoS consensus mechanism that is a suitable solution for the scalability problem.

Table 4: Comparison of off-chain scalability solutions

Solution	Mechanism	Scalability measurements				Advantages	Disadvantages
		Throughput	Latency	Storage	Bandwidth		
Lightening Network [39]	Payment channel	High↑	Low↓	Low↓	Low↓	- Instant transactions - Lower fees	- Less secure - Supports only bitcoin micropayments - Needs the users to be online at the same time and follow same payment route
Raiden [40]	Payment channel	High↑	Low↓	Low↓	Low↓	- Fast and free transactions - Enables multi-hop transfers - Supports general purpose transactions	- Implies more complexity because it allows multi-hop transfers
µRaiden [41]	Payment channel	High↑	Low↓	Low↓	Low↓	- Fast and free ERC20 token - Supports general purpose transactions	- Does not allow multi-hop transfers
Trinity [42]	Payment channel	High↑	Low↓	Low↓	Low↓	- Real-time payment - Low-cost transactions - Privacy protection	- Supports only payment transactions
Rapido [43]	Payment channel	High↑	Low↓	Low↓	Low↓	- Avoids overload issue - Prevents privacy leaking - Mitigates the skewness and congestion issue	- Discourages individual donation
Plasma [44]	Sidechain	High↑	Low↓	Low↓	--	- Hierarchical structure - Reduces the congestion of the main blockchain - Fast and low-cost transactions - No need to users be online at the same time	- Complicated to be implement - Long waiting time for transferring assets to the main chain
SPRF [45]	Sidechain	--	--	Low↓	--	- Secure and also computationally efficient - Applicable to the existing blockchain without any modifications	--
Atomic-swap solutions [46]	Inter-chain	--	--	--	--	- Solve interoperability between different blockchains	- Work under specific situations

II) Practical Byzantine Fault Tolerance

Practical Byzantine Fault Tolerance (PBFT) [49] is a variation of the vote/BFT-based consensus algorithm, in which the nodes reach consensus using a collective decision-making strategy even if some nodes withhold responding or respond incorrectly. PBFT operates in successive rounds called views. Each view has a primary node called leader and other nodes are referred to as backup nodes. The leader is changed in every view. PBFT consensus rounds consist of three phases: pre-prepare, prepare and commit.

Fig. 9 [116] shows an illustration of the PBFT consensus algorithm. As is shown, in the pre-prepare phase, the leader multicasts the next record (block) to the backup nodes. Then, in the prepare phase, after receipt of the pre-prepare message, the backup nodes validate its veracity and multicast a prepare message to all the other nodes in the consensus group. After that, in the commit phase, upon receiving prepare messages from more than two-thirds of all the nodes, each backup node multicasts a commit message to the consensus group and then waits for more than two-thirds of commit messages, to ensure that the majority of nodes have come to the same decision. Consequently, all the honest nodes agree unanimously on the valid record. Although PBFT is energy efficient and increases the transaction rate, it suffers from high communication overhead. Therefore, it is not scalable enough to be used in public networks and thus is only applicable to private and permissioned networks. Moreover, PBFT mechanisms are vulnerable to Sybil attacks where an adversary takes over the network by creating multiple fake identities for malicious purposes. Hence, the PBFT mechanisms are usually used in combination with other mechanisms.

For example, Hyperledger Fabric [90], an open-source blockchain framework for developing blockchain-based applications, has utilized a permissioned version of PBFT. In addition, Zilliqa [66] is a high-throughput blockchain that uses PBFT for consensus-making together with PoW for establishing identities. Tendermint [117] is also a consensus protocol that merges PBFT with DPoS to bring PBFT to a public blockchain.

III) Federated Byzantine Agreement

Federated Byzantine Agreement (FBA) [50] is another BFT-based consensus mechanism operating based on "quorum" and "quorum slice" concepts. The quorum refers to the nodes that should reach a consensus on the information that is to be stored in the blockchain. The quorum consists of individual quorum slices that are subsets of the quorum nodes. The transactions are confirmed only if a required number of the quorum slices agree on it. Hence, the FBA data ledger can be updated without requiring all the nodes to agree, resulting in network scalability and fast transaction with low cost.

Stellar [50] and Ripple [84] are two main cryptocurrencies using the FBA consensus mechanism. Steller has implemented an enhancement of FBA. It provides an open membership so that anyone can join the network or even be a validator without the need to be verified ahead of time. In addition, in the Stellar network, users can determine which quorum slice they trust. On the other side, Ripple has a close membership and only pre-selected validators vote on the veracity of the transactions. Therefore, Steller is more decentralized than Ripple.

IV) delegated Byzantine Fault Tolerance

delegated Byzantine Fault Tolerance (dBFT) consensus algorithm [51] was first introduced by Neo blockchain [52], a smart contract platform that is often referred to as "Ethereum of China". Generally speaking, there are three types of nodes in dBFT, called speaker, delegate and common node. The common nodes are the ordinary token holders that vote to elect delegates. Delegates form a consensus group for BFT consensus. Then, a speaker is randomly chosen among the delegates. The transactions created by the common nodes are received by the speaker node. The speaker node validates the received transactions and then creates a block containing a number of valid transactions. After that, the speaker multicasts the new block to the delegates. Upon receiving the new block, the delegates validate it separately and respond to the speaker (same as the process used in the PBFT). If more than two-thirds of the delegates confirm the new block, it will be added to the chain. The Neo employing dBFT provides a high throughput, making it applicable to large-scale commercial applications. Neo was designed to digitize assets using smart contracts and enables users to trade their digitized assets by two types of tokens namely Neo and GAS. dBFT is also used by other blockchains such as ONT [118]. A major disadvantage of dBFT is that the delegates require to provide a real identity to be elected during the voting process.

V) Proof of Authority

Proof of Authority (POA) [53], [54] is an alternation of PoS that uses identity as a stack. In the PoA protocol, a number of trusted nodes called validators are responsible for validating the transactions. A leader is randomly selected from the set of all validators to add a new block to the chain. Any leader that does not perform appropriately will be voted out and replaced by other validators.

POA leads to a scalable and high-throughput blockchain but due to its identity-based and centralized nature, it is more applicable in the private blockchains than public ones. POA Network [119] is the first public Ethereum-based platform that has employed the PoA consensus mechanism. It provides an open-source framework for smart contracts. All validators within PoA Network are licensed by United States notaries, and their identities reference a public notary database.

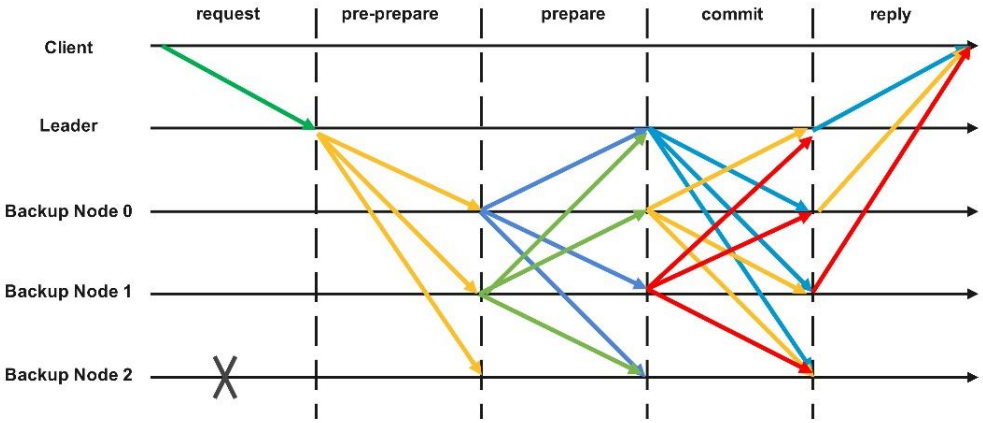


Fig. 9. Practical Byzantine Fault Tolerance (PBFT) protocol [116].

VeChain [120] is another blockchain project using PoA. In the VeChain network, the validators are called Authority MasterNodes (AMs) and to be an AM, users need to submit identifying information to the VeChain Foundation. VeChain is in an effort to enhance its PoA to provide a more randomized and distributed block-creating mechanism.

VI) Algorand protocol

Algorand [55] is a blockchain-based cryptocurrency that utilizes a new Byzantine Agreement (BA) called BA★ allowing users to achieve consensus on the next set of transactions with low latency. It is also able to scale consensus to millions of users. To achieve scalability, Algorand selects randomly a few representatives from the entire set of users. Representatives form a committee responsible for confirming transactions. to prevent Sybil attacks, Algorand assigns a weight to each one of the users based on the stack they own and as long as a weighted fraction of the users are honest, the consensus is guaranteed by BA★ protocol.

In addition, Algorand utilizes a novel mechanism based on Verifiable Random Functions (VRF) enabling it to choose committee members through a random and private way. Meaning that, by calculating a VRF of their private key and some public information in the blockchain, the participating users are able to individually specify whether they are selected to be on the committee or not. This prevents attackers to recognize the committee members ahead of time and plan a target on them. After computing VRF, only committee members have the right to propose a new block, thus they propagate the proposed block along with the VRF output which proves that the account is a committee member.

To achieve consensus using BA★ and ensure that all the nodes have the same view of the blockchain, the nodes will confirm the signature of the message containing proposed blocks and then, using the VRF proof, validate whether the proposer is a committee member or not. Next, through a cryptographic sortition, each node will compare the hash of the messages received from the committee members to identify the lowest one and then will only propagate the block proposal with the lowest VRF hash. Consensus Process consists of several interactive steps and continues until when a proposed block receives enough votes from weighted committee members. Therefore, it should be said that BA★ exploits a Pure PoS.

To sum up, Algorand is a fast, scalable and secure cryptocurrency. Moreover, it supports smart contracts and all kinds of financial transactions and programs, however, it still has not been adopted widely in the world of cryptocurrencies.

Comparison of Scalable Consensus Mechanisms

Table 5 provides a plain comparison between some common characteristics of the discussed scalable consensus mechanisms to help better understand their contributions.

O. DAG-based scalability solutions

Directed Acrylic Graph (DAG)-based data ledgers are an alternative to blockchain-based data ledgers, being a potential solution to address scalability issues. A DAG-based ledger is a network of individual transactions in which there are no blocks of the transactions and competition for appending new blocks, thus the confirmation time of the transactions is not bound to the interval of the blocks in the blockchain and transactions are processed independently. Consequently, the throughput of the data ledger is improved notably. Furthermore, DAGs alleviate transaction fees because the DAG-based consensus algorithms are simpler than the

Table 5: Comparison of scalable consensus mechanisms

Solution	Technique	Existing Projects	Scalability measurements				Advantages	Disadvantages
			Throughput	Latency	Storage	Bandwidth		
Delegated Proof of Stack (DPoS) [47], [48]	A variation of PoS combined with voting mechanism	EOS [85], BitShares [48], Steemite [112], Ark [113], Lisk [115], Cardano [114]	High↑	Low↓	--	--	- Scalable and fast - Better distribution of rewards - Energy efficient	- Partially centralized - Less secure against 51% attack
Practical Byzantine Fault Tolerance (PBFT) [49]	BFT- based	Hyperledger Fabric [90], Zilliqa [66], Tendermint [117]	High↑	Low↓	--	High↑	- High transaction rate - Energy efficient	- High communication overhead
Federated Byzantine Agreement (FBA) [50]	BFT- based and quorum-based	Stellar [50], Ripple [84]	High↑	Low↓	--	--	- Fast transaction - Low cost	--
delegated Byzantine Fault Tolerance (dBFT) [51], [52]	BFT-based	Neo [52], ONT [118]	High↑	Low↓	--	--	- High transaction throughput - Low latency - Energy efficient	- There is no anonymity and delegates need a real identity
Proof of Authority (PoA) [53], [54]	An alternation of PoS (identity as stack)	PoA Network [119], VeChain [120]	High↑	Low↓	--	--	- Scalable and high-throughput - Energy efficient	- Has an identity-base and centralized nature - More suitable for private blockchains
Algorand [55]	Pure PoS	--	High↑	Low↓	--	--	- High transaction throughput, par with large payment and financial networks - Scalable to many users - Low latency - Low transaction fee - Secure against DOS and Sybil attacks	- Not adopted widely

ones used by traditional blockchains. Generally, DAG-based solutions focus on both the data layer and consensus layer, to improve scalability. To do so, they adopt a novel data structure based on DAG and employ a consensus mechanism convenient to it.

In the following, some of the prominent DAG-based scalability solutions are discussed. In addition to the works that will be explained in this section, there are several other DAG-based data ledgers in the literature such as ByteBall [121], Spectre [122], GraphChain [123], Phantom [124], CDAG [125], Conflux [126], Dexon [127], Teegraph [128] and so on. More details about these and other DAG-based data ledgers are available in [98].

I) IOTA

IOTA [56] is the most popular DAG-based data ledger that has initially been designed for the Internet of Things (IoT) industry. IOTA is based on a data structure named Tangle. Tangle [76] is a particular type of DAG, made up of transactions that are connected by edges. Each edge from transaction A to transaction B indicates that transaction A validates and approves transaction B. An illustration of a Tangle had been shown before in Fig. 2.

In IOTA, before sending out a new transaction, users need to solve a simplified PoW problem and then validate two previous transactions simultaneously. Each user technically acts as a miner that mines the previous transactions to be able to send new transactions. Therefore, there is no transaction fee in IOTA. In addition, since the previously added transactions are validated by the new transactions, the more transactions are created by the users, the more transactions are confirmed per second. Subsequently, the throughput and confirmation time of the transactions are improved. Another advantage of the IOTA is that it provides security against quantum computers as it uses hash-based signatures rather than elliptic curve cryptography. On the other side, one of the disadvantages of IOTA is that it lacks smart

contracts, thus developing DApps on the IOTA is almost impossible.

II) Nano

Nano [57] uses a novel form of DAG namely Block-lattice. The Block-lattice architecture is a hybrid between blockchain and DAG and is made up of account-based blockchains. This means that each user holds a separate blockchain for each account, that represents the transaction history of that account. The participating users can only control and update their own individual blockchain with their private keys. Meanwhile, they update blockchains owned by other accounts asynchronously, rather than forming an agreement on a shared data ledger. Every transfer of nano coins requires two separate transactions/blocks, a send transaction deducting the amount from the sender's balance and a receives transaction adding the amount to the receiver's balance (see Fig. 10 [57]). The send transaction is signed by the sender account and is stored on the account chain of the sender while the receive transaction is signed by the receiver account and is stored on the account chain of the receiver. Furthermore, each transaction contains the current balance of its owner account.

Nano utilizes a variation of Delegated Proof-of-Stack (DPoS) consensus mechanism called Open Representative Voting (ORV) under which account holders can choose a representative to vote on their behalf regarding the validity of the transactions, even when the delegating account is itself offline. One drawback of Nano is that it is prone to spam attacks, meaning that it can be flooded with spam transactions, causing some valid transactions to be obstructed and network nodes to be out of sync.

III) DLattice

DLattice [58] is a permission-less blockchain with a double-DAG architecture under which Dlattice provides data protection and tokenization. DLattice has a double-DAG structure because each account has its own account-DAG and all account-DAGs form a greater DAG namely Node-DAG. Node-DAG organizes all the Account-DAG in the form of a Merkle Patricia Tree (MPT) using a Genesis header. Each Account-DAG structure consists of a token-chain and a data-tree. The token-chain is a unidirectional chain that records the income and expenditure history of the digital assets sent by the account, whereas, data-tree is a Red-Black Merkle Tree [129] combined with token-chain, that stores the digital fingerprint of the data asset and corresponding access control permissions. DLattice isolates transaction processing within each Account-DAG. Therefore, the transactions of the accounts can be processed in parallel resulting in fast transactions with minimal overhead. Instead of executing consensus at a fixed interval, DLattice uses a new DPoS-BA-DAG (PANDA) protocol to reach a low latency consensus among users only when the forks are observed. There are also some issues against DLattice. For example, it does not support smart contracts and also it is prone to DDoS attacks focusing on flooding false transactions and attacking smart contracts.

IV) Hashgraph

Hashgraph [59] is a DAG-based consensus algorithm based on a gossip protocol. In gossip protocol, each participating node randomly communicates with other nodes in the network to inform them about all the information it has, until the whole network is aware of all the transactions that have been processed so far.

In fact, nodes gossip about gossip. This means that they not only gossip about transactions but also gossip about the information that they have received from other nodes. Each member in the network maintains a separate chain to record the history of all the gossip events during which it receives some information. As it can be seen in Fig. 11 [59], the network members will eventually build a full history and create collaboratively a Hashgraph of all the gossip events. Then, each event in the Hashgraph is validated during a conventional Byzantine Fault Tolerance (BFT) consensus procedure. Hashgraph also enables visual voting, meaning that if two nodes have the same Hashgraph, rather than sending a vote message they can calculate each other's vote. Hence, the Hashgraph algorithm has very little communications overhead. Totally, Hashgraph-based communication patterns result in fast convergence of the information at all the nodes. All these features make Hashgraph a fair, fast and Byzantine Fault Tolerant solution, however, one drawback of the Hashgraph is that it is not secured against Sybil attacks thus it is more suitable for a permissioned network.

V) JHdag

JHdag [60], [61] is a PoW-based consensus mechanism that is designed on a novel DAG structure under which the network members can reach consensus at a large scale. In this structure, each block only contains one transaction in order to save network bandwidth when broadcasting blocks. Furthermore, the PoW puzzle is simplified to scale up processing capacity. Additionally, a mempool transaction assignment mechanism is designed based on the DAG structure to reduce the probability of processing a transaction by multiple miners, and hence reduce the waste of the capacity. To reach a consensus, a Nakamoto chain is embedded in a DAG structure that is strongly connected and incorporates miner information. There exist two types of blocks: regular blocks for carrying transactions and milestone blocks for making decisions. Blocks on the embedded Nakamoto chain are milestones

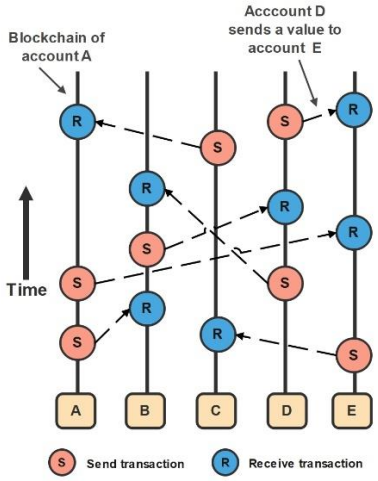


Fig. 10: Block-Lattice structure in Nano blockchain [57].

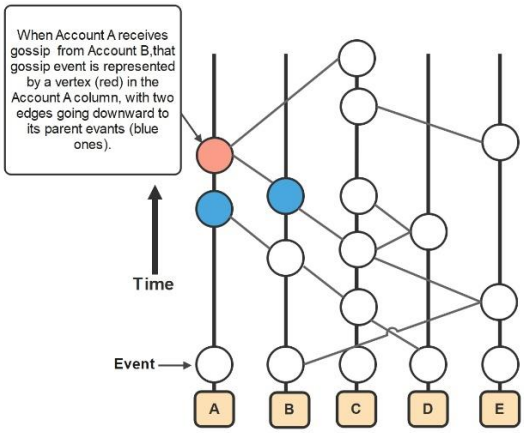


Fig. 11: Hashgraph, a full history of all gossip events [59].

and are harder to mine than regular blocks and each milestone block can verify multiple regular blocks.

VI) TrustChain

TrustChain [62] is a scalable, tamper-proof and Sybil-resistant blockchain, working based on the notion of trust between nodes. Each node has its own temporally ordered chain of blocks containing transactions that it participates in. Each block in TrustChain includes at most one transaction signed by both transacting parties and is linked back to the last block in the chain of both participating nodes using hash values. Accordingly, each block has two incoming and two outcoming pointers. As a result, transactions are arranged in a trusted and tamper-proof manner and form a DAG structure capable of creating trusted transactions without the need for any central authority or global consensus. In addition, to avoid Sybil attacks, TrustChain offers a new algorithm named NetFlow to calculate the trustworthiness of the network nodes according to the TrustChain graph. Indeed, the trustworthiness score which is assigned to each node in the graph determines if that node can contribute back the resources that it needs. Consequently, the network is affected only by the nodes with a positive score. Having said that, the transaction throughput of Trustchain is approximately 210 transactions per second which is not very high compared to centralized payment systems and some of the other scalable blockchains.

Comparison of DAG-based Scalability Solutions

In Table 6, a comparison of DAG-based solutions is presented. The promise of DAG-based solutions is to parallelize transaction processing whereby a high scalability is achieved in many aspects e.g. throughput and latency. For this purpose, they alter the data structure in the data layer of traditional blockchain and usually run a different algorithm in the consensus layer. To provide a clear vision, Table 6 summarizes some of the characteristics of the described DAG-based solutions, such as their data structure and consensus mechanism and also specifies their influence on the scalability measurements.

P. Horizontal Scalability Through Sharding

Sharding is a horizontal scaling solution in which adding more nodes to the network increases system performance. It refers to the techniques that partition blockchain nodes into subsets called shards. The workload of the blockchain is also distributed among the shards acting in parallel, leading to a high-performance and high-throughput blockchain. In addition to horizontal scaling, there are few works towards vertical scaling of the nodes, e.g. Ostraka [130], in which instead of partitioning nodes in a blockchain system into shards, each node itself is sharded into multiple Node-Shards.

Fig. 12 illustrates an example of sharding architecture, where the transactions are distributed among multiple shards and processed in parallel. In the following, some of the sharding-based blockchains and their properties are briefly explained.

I) Elastico

Elastico [63] is the first sharding-based protocol for permission-less blockchains in which nodes have no pre-published identities. Elastico breaks up the network into multiple committees each of which handles a disjoint set of transactions. A “consensus committee” is responsible for combining agreed transaction sets of other committees. Elastico uses a PoW mechanism to establish identities and map them randomly to the committees and also uses a PBFT mechanism to reach consensus within each committee. The protocol proceeds in epochs. At the beginning of each epoch, the identities are re-established and the committees are reconstructed.

Table 6: Comparison of DAG-based scalability solutions

Solution	Structure	Mechanism	Consensus	Scalability measurements				Advantages	Disadvantages
				Throughput	Latency	Storage	Bandwidth		
IOTA [56]	Tangle	--	Cumulative weights of all transactions that directly or indirectly approve the transaction	High↑	Low↓	--	--	- Fast confirmation - No transaction cost - High throughput and low latency - Quantum resistant	- Does not support smart contracts yet
Nano [57]	Block-Lattice	Independent nonshared blockchains	ORV	High↑	Low↓	--	--	- Fast confirmation - High throughput and low latency - Fee-less	- Prone to spam attack
DLattice [58]	Double-DAG	--	PANDA	High↑	Low↓	--	--	- Fast confirmation - No transaction cost - High throughput and low latency - Provides data tokenization	- Prone to DDoS attack - Does not support smart contracts
Hashgraph [59]	Hashgraph built of all the gossip events	Gossip about Gossip/ event-based	BFT	High↑	Low↓	--	Low↓	- Fair, fast and Byzantine Fault Tolerant - Visual voting	- Not secured against Sybil attacks
JHdag [60], [61]	Embedded Nakamoto chain inside DAG structure	Flexible-PoW	PoW	High↑	Low↓	--	Low↓	- Little or no transaction fee - Save bandwidth - Concentration of mining power within mining pool - Reduces waste of capacity	--
TrustChain [62]	A chain of trusted transactions	Building trust between individuals	consensus is reached among transacting users without needing any global consensus	High↑	Low↓	--	--	- Sybil-resistant - Removes the requirement for global consensus	- not very high transaction throughput compared to centralized payment systems

Elastico can scale linearly with the number of participating nodes and it is efficient in terms of communication overhead. In addition, the network topology between honest nodes is connected and the communication channel is partially synchronous. In terms of resiliency, Elastico can tolerate malicious users controlling up to one-fourth fraction of the total computational power in the network, while resiliency for each committee is one-third of malicious processors. That is not good enough.

Although Elastico can improve scalability measurements such as throughput and latency considerably, it has some shortcomings as follows. First, it is not storage efficient as it needs all users to store the entire data ledger. Second, in order to reduce the communication overhead of running PBFT consensus, it needs to choose a small committee size that leads to a high failure probability. Third, Elastico does not guarantee the atomicity of the cross-shard transactions. Finally, the epoch randomness used by Elastico for establishing identities and formation of the committee is not fully bias-resistant and might be biased by malicious users.

II) OmniLedger

OmniLedger [64] is a secure and permission-less blockchain that provides scalability via sharding. To securely assign nodes to the shards, OmniLedger implements sharding using the Randhound protocol that provides a bias-resistant decentralized randomness. In order to process inter-shard transactions, shards run ByzCoinX, an enhancement of PBFT-based consensus in ByzCoin [131], that improves performance and robustnes

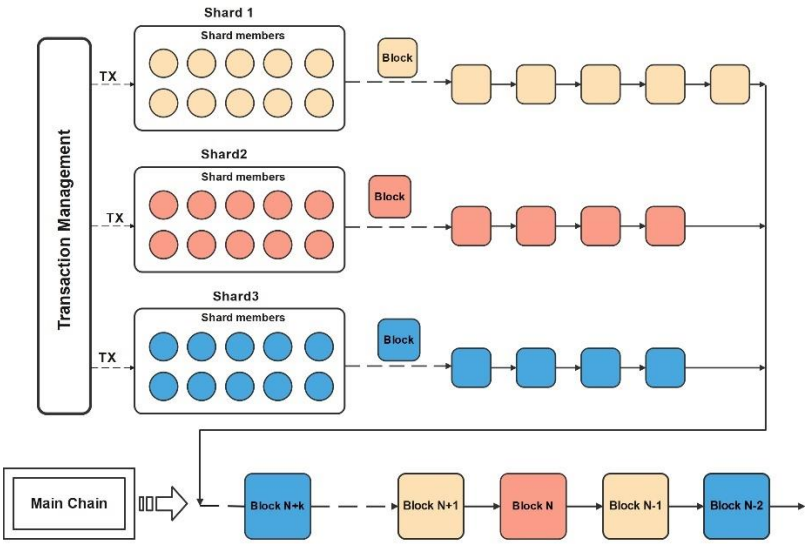


Fig. 12: Architecture of sharding-based blockchain.

against DoS attacks. Like Algorand [55], OmniLedger uses VRF and cryptographic sortition to pick a subset of the validators based on some per-validator weight functions. OmniLedger has also proposed the Atomix protocol that ensures atomically processing of cross-shard transactions. To optimize storage and reduce update overhead, OmniLedger uses the state blocks that provide checkpoints for the data ledger. In addition, OmniLedger provides low latency for low-value transactions using two-step trust-but-verify processing and also shows a latency of seconds for typical transactions.

Omniledger is a full-decentralization blockchain and has no single points of failure. In OmniLedger, the throughput increases almost linearly as the number of participating nodes increases. Furthermore, OmniLedger offers a throughput advantage par with centralized payment systems such as Visa, without compromising security or decentralization. It also is able to handle Visa-level workload.

On the other hand, similar to Elastico [63], OmniLedger can only tolerate up to one-fourth of malicious nodes in the entire network, and up to one-third of malicious nodes in each committee. In addition, per each confirmed block, the OmniLedger protocol needs to gossip multiple messages to all the nodes in the network. Another drawback of OmniLedger is that it needs the client to be an active participant in cross-shard transactions which is an inconvenient assumption for lightweight clients.

III) RapidChain

RapidChain [65] is the first one-third resilient sharding-based blockchain protocol scaling public blockchains via full sharding of computation, communication and storage. Meaning that, in addition to parallelizing transaction processing, the data ledger is also divided into partitions each of which is stored by one of the committees. Furthermore, to provide sharding of communication, RapidChain uses the Kademlia routing algorithm for committee-to-committee communication and cross-shard transaction processing. On the other hand, to process intra-committee transactions, each committee chooses a leader based on epoch randomness. The leader forms a new block and creates the block header and then propagates the block header using IDA-gossip protocol. Finally, the committee runs a synchronous BFT consensus protocol to make a consensus on the header of the block. Consequently, RapidChain needs only a sublinear number of bits to be exchanged per transaction.

To prevent Sybil attacks, RapidChain needs the nodes wanting to join the network to solve a PoW puzzle. All nodes solve PoW offline to avoid any interruptions in the protocol execution. In addition, to prevent a slowly-adaptive adversary from compromising one or more committees, RapidChain runs a reconfiguration protocol built on the Cuckoo rule [132] between epochs, without regenerating all the committees. In Rapidchain, total resiliency and committee resiliency are improved to one-third and one-two respectively. In addition, Rapidchain shows a much higher throughput and better latency than Elastico [63] and Omniledger [64].

IV) Zilliqa

Zilliqa [66] is a public blockchain platform designed to increase the transaction rate using sharding that enables parallel processing of transactions on multiple shards. Zilliqa also provides a smart contract platform and innovates a special-purpose smart contract language that follows a dataflow programming style that facilitates parallelizing of large-scale computation. Zilliqa uses PoW to establish node identities and prevent Sybil attacks. To reach consensus, Zilliqa uses an evolution of the PBFT

algorithm that is inspired from ByzCoin [131] and replaces the Message Authentication Code (MAC) used in the classical PBFT with a digital signature to lessen communication overhead to O(n) and also employs EC-Schnorr multisigniture to aggregate several signatures into an O(1)-size multisignature.

In addition, inspired from Bitcoin-NG [133], Zilliqa adopts two types of blocks: Transaction Blocks (TX-Block) and Directory Service Blocks (DS-Block). TX-Block includes the transactions sent by the users, whereas DS-Block includes metadata about the miners who participate in the consensus protocol. A TX-block will be finalized if it contains an EC-Schnorr multisignature by more than two-thirds of the miners.

Zilliqa reaches a great transaction rate, about a thousand times of Ethereum, although it shows the same local and global resiliency as Elastico [63] and Omniledger [64]. Zilliqa also suffers from some shortcomings. First, it does not provide storage sharding (state sharding). Meaning that all full nodes need to store and receive all the blocks and transactions, resulting in a high storage requirement. Second, Zilliqa is vulnerable to single-shard takeover attacks since it relies on PoW as a randomness generation mechanism.

V) Harmony

Harmony [67] is a fully scalable blockchain that similar to Rapidchain provides full sharding for transactions, communication and storage. To prevent Sybil attack and select validators, Harmony uses PoS rather than PoW, making it energy efficient. Consensus is reached using a new algorithm called Fast Byzantine Fault Tolerance (FBFT) which is linearly scalable in terms of communication complexity and is at least 50% faster than PBFT.

In FBFT, the leader executes a multi-signature signing process to collect the votes of the validators. Indeed, the sharding process is based on PoS and voting shares. Stackers gain voting shares proportional to their stack amount and then are randomly assigned to the shards. Like RapidChain [65], Harmony uses a cuckoo-based mechanism for resharding.

Harmony has proposed a unique algorithm for randomness generation by a combination of VRF and Verifiable Delay Function (VDF) that is unpredictable, unbiaseable, verifiable and scalable. In addition, Harmony uses RaptorQ function code to speed up the block propagation process within shards. It also uses an atomic locking mechanism to guarantee the consistency of the cross-shard transactions and adopts Kademlia as a routing mechanism for cross-shard communication and reducing communication overhead. The local (committee) and global resiliency of Harmony is the same as Elastico [63], omniledger [64] and Zilliqa [66].

VI) Monoxide

Monoxide [68] has been designed to scale out blockchain systems linearly without sacrificing security and decentralization. Towards this purpose, it divides the blockchain network into multiple independent and parallel instances called “consensus zones” and partitions the workload of computation, communication, storage and memory (for state representation) in the consensus zones. Each consensus zone has its own chain of blocks and runs the consensus process independently with minimized communication. In Monoxide, the blocks are created by miners. Although Monoxide uses PoW for mining blocks, its technique is impertinent to the consensus mechanism used per zone. To prevent attackers from controlling more than 50% of mining power in a single zone, Monoxide proposes a Chu-ko-nu mining mechanism that allows a miner to create multiple blocks in the different zones with one PoW solution. Consequently, the mining power is dispersed into multiple zones. since a PoW solution in Chu-ko-nu mining is more productive of blocks, it is more energy-efficient than traditional PoW. In addition, Monoxide uses eventual atomicity to ensure the atomicity of cross-zone transactions. In contrast to two-phase commit protocols that serialize the transactions, eventual atomicity has no additional delay and overhead.

VII) FleetChain

FleetChain [69] is a scalable and responsive blockchain with optimal sharding focusing on the intra-shard consensus and cross-shard transactions. To achieve an efficient intra-shard consensus, FleatChain proposes a Leader-Stable Fast Byzantine Fault Tolerance (FBFT) protocol that adopts a multi-signature schema to reduce message size during voting, combined with pipeline technology to enhance processing efficiency. Furthermore, for cross-shard transactions, a Responsive Sharding Transaction Processing (RSTP) protocol has been introduced that depends on the classical two-phase commit (2PC) protocol in which transaction inputs are locked/unlocked. Albeit unlike Omniledger protocol [64] where a client is considered as coordinator of the cross-shard transactions, in Fleetchain output shard leader operates as coordinator.

For intra-shard consensus, FleatChain utilizes a robust t out of u Multi-Signature Protocol with public key aggregation using Proof-of-Possession (PoP), shortly referred to as (t, u)-MSP-PoP, while uses (t, u)-AMSP-PoP (Aggregated Multi-Signature) for cross-shard transactions. To sum up, Fleetchain is scalable from the perspectives of computation (i.e., transaction throughput and latency), communication and storage, and its

scalability factor is $O(n/\log n)$ where n represents the network size.

Comparison of Horizontal Scalability/ Sharding Solutions

Table 7 and Table 8 provide a summarized comparison of the described sharding-based blockchains for a better understanding of their techniques and key features.

Future Directions and Open Issues

Although numerous researches addressing scalability challenges have been proposed in recent years, there are still some issues that were not resolved in the best way possible and were left open for future works. In this section, the open issues and future research directions of each category of scalability solutions are discussed separately.

Q. On-Chain Scalability

Secure blockchain pruning: large blockchain size leads to centralization problems due to limited storage capacity. An approach for reducing blockchain size is to remove non-critical and stale blockchain information to free up storage space on the nodes. Although a number of works [134] focusing on blockchain pruning have recently been proposed, there is still an outstanding question that needs to be answered: what data at what time must be removed so that security would not be compromised.

Blockchain data query: decentralization and data distribution in the blockchain lead to inconvenience for querying required data. As the blockchain grows, processing the various queries such as single, range and condition ones among a large amount of data, goes through performance and bandwidth issues. Hence, providing an efficient solution for querying blockchain data is an open issue that has not received enough attention in the literature.

R. Off-Chain Scalability

Future work can direct off-chain blockchains towards further off-chain computation techniques and conduct hybrid off-chaining mechanisms.

S. Scalable Consensus Mechanisms

Novel proof-based consensus mechanisms: Most scalable consensus mechanisms in the literature are based on voting, while there exist not many proof-based scalable ones. Therefore, designing secure and low latency proof-based algorithms is a topic that needs to be studied more. For example, an idea is to develop protocols that adopt non-transferable incentives such as reputation or familiarity, in which mining difficulty can be dynamically controlled.

Multi-block consensus mechanisms: redesigning consensus protocols so that they will be able to reach consensus on multiple blocks can improve throughput considerably.

T. DAG-Based Scalability

Trade-off: existing DAG-based systems failed to make a trade-off between multiple factors. For example, IOTA [56] and GraphChain [123] enhance performance and scalability while compromising security and consistency. On the other side, some DAG-based data ledgers e.g. Prism [135] and OHIE [136] provide strict consistency at the cost of scalability and performance. Hence, designing a DAG-based solution that can reach a balance between various metrics is still a challenging issue.

Supporting off-chain transactions: redesigning DAG-based systems for supporting off-chain transactions is an interesting direction for future work, which take advantage of both off-chain and DAG-based solutions.

System setup: setup configuration defines all the specific needs that must be available at the onset of the protocol to each participating node. Some existing DAG-based systems [56], [123], [137], [138] rely on the genesis block, whereas some others [57]-[59], [139], [140] initialize multiple parallel chains simultaneously, however, system setup using these parallel chains is unclear. Therefore, adopting a novel and transparent system setup can be considered as a future work.

U. Horizontal Scalability Through Sharding

Cross-shard transaction: cross-shard transactions lead to a lot of communication overhead and also reduce system performance and increase transaction confirmation time. Therefore, assigning transactions to different shards in a way that the cross-shard transactions be minimized is still an open issue. In this regard, the authors in [141] proposed a new sharding paradigm with optimal transaction placement, called OptChain. Furthermore, in [142] a new scalable permissioned blockchain named “Sharper” has been introduced that shards the transaction processing through clustering network nodes. Using two decentralized flattened consensus protocols, Sharper handles cross-shards transactions more efficiently. However, there is still a need for more efficient protocols for processing cross-shard transactions to reduce confirmation latency.

Resharding: resharding process is a challenging issue in sharding-based blockchains since it needs reshuffling the network which leads to huge data migration. SSChain [143] is the first public blockchain that provides full sharding with no reshuffling process and data migration.

Adaptive malicious attackers: resharding process is performed to prevent malicious users from overtaking a shard by corrupting the members of that shard during protocol epochs. Securing committee members against both slowly adaptive and fully adaptive attackers is a crucial problem that must be taken into consideration.

Table 7: Comparison of horizontal scalability solutions through sharding

Solution	Transaction Model	Identity Setup/ Committee Formation Mechanism	Intra-Consensus Mechanism	Cross-Shard Transactions Mechanism	Smart Contract	Total Resiliency	Committee Resiliency
Elastico [63]	UTXO	PoW	PBFT	Not supported	×	$1/4$	$1/3$
OmniLedger [64]	UTXO	RandHound	ByzCoinX	Sync, Lock/Unlock (AtomiX)	×	$1/4$	$1/3$
RapidChain [65]	UTXO	Offline PoW	Synchronous BFT	Sync, lock/Unlock	×	$1/3$	$1/2$
Zilliqa [66]	Account	PoW	An evolution of PBFT	not Supported	✓	$1/4$	$1/3$
Harmony [67]	Account	PoS	FBFT	Sync, Lock/Unlock	✓	$1/4$	$1/3$
Monoxide [68]	Account	Consensus zones (partitioning based on users address)	PoW (Chu-ko-nu mining)	Async, Lock-free (Eventual atomicity)	×	$1/2$	$1/2$
FleetChain [69]	UTXO	Proof of Possession (PoP) combined with PoW	Leader-Stable FBFT	Sync, lock/Unlock	✓	N/A	1/3

Table 8: Comparison of horizontal scalability solutions through sharding (tps: transactions per second, s: second, n: network size, m: committee size)

Solution	Sharding Components			Throughput *	Latency *	Communication Complexity
	Computation	Communication	Storage			
Elastico [63]	✓	×	×	48ktps	<900s	$O(m^2 + n)$
OmniLedger [64]	✓	×	×	28.8ktps	~100s	$O(log_2^m + n)$
RapidChain [65]	✓	✓	✓	128ktps	70s	$O(m^2 + mlog_2^n)$
Zilliqa [66]	✓	×	×	N/A	N/A	$O(n)$
Harmony [67]	✓	✓	✓	N/A	N/A	$O(log_2^n)$
Monoxide [68]	✓	✓	✓	1.23~2.56Mtps	23s	$O(m + n)$
FleetChain [69]	✓	✓	✓	N/A	N/A	$O(n/m)$

* The indicated throughputs and latencies are according to the evaluation of some sharding-based mechanisms conducted by Yu et al. in [97]. For more information about evaluation conditions, refer to Table 3 in [97].

An interesting idea is to incorporate the sharding process with machine learning algorithms to analyze the behavior patterns of users on the network and detect malicious users.

Conclusions

Scalability is the most important challenge to blockchain mass adoption.

This paper focuses on the blockchain scalability issue and reviews some related works in the literature dealing with it.

To do so, the scalability solutions are firstly classified into five categories including on-chain, off-chain, scalable consensus mechanism, DAG-based scalability, and sharding solutions.

Then, the key properties of these solutions along with their advantages and disadvantages are discussed to reveal their main contributions.

In addition, the discussed works are compared in terms of scalability improvements such as throughput, latency, storage and bandwidth. Finally, the future trends and open issues expected to be investigated through future works are discussed.

This paper provides a deep understanding of existing scalability solutions as well as the issues and challenges they deal with. Hence, it inspires novel ideas for more scalable and efficient blockchains in the future.

Credit Authorship Contribution Statement

Alemeh Matani: Writing- original draft, Investigation, Conceptualization. Amir Sahafi: Writing- review & editing. Ali Broumandnia: Writing- review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is affiliated with Islamic Azad University, South Tehran Branch, Tehran, Iran. Meanwhile, it did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abbreviations

TPS	Transactions Per Second
DAG	Directed Acrylic Graph
UTXO	Unspent Transactions Output
PKI	Public Key Infrastructure
MPT	Merkle Patricia Trie
SPV	Simplified Payment Verification
ECDSA	Elliptic Curve Digital Signature Algorithm
EdDSA	Edwards-curve Digital Signature Algorithm
BRS	Borromean Ring Signature
OTS	One-Time ring Signature
PoW	Proof of Work
PoS	Proof of Stack
P2P	Peer-to-Peer
BFT	Byzantine Fault Tolerance
PBFT	Practical Byzantine Fault Tolerance
VM	Virtual Machine
EVM	Ethereum Virtual Machine
DApps	Decentralized Applications
RQ	Research Question
MAST	Merkelized Abstract Syntax Tree
AST	Abstract Syntax Tree
AB-M	Adaptive Balanced Merkle
VDP	Value Distributing Problem
DHTC	Distributed Hashed Timelock Contracts
SPRF	Smart Program Runner Framework
PoA	Proof of Authority
PoET	Proof of Elapsed Time
PoC	Proof of Capacity
PoI	Proof of Importance
PoB	Proof of Burn
DPoS	Delegated Proof of Stake
dBFT	delegated Byzantine Fault Tolerance
FBA	Federated Byzantine Agreement
BA	Byzantine Agreement
VRF	Verifiable Random Functions
FBFT	Fast Byzantine Fault Tolerance
VDF	Verifiable Delay Function
RSTP	Responsive Sharding Transaction Processing
PoP	Proof-of-Possession

References

[1] Y. Chen, C. Bellavitis, "Blockchain disruption and decentralized finance: The rise of decentralized business models," J. Bus. Ventur. Insights, 13: e00151, 2020.

[2] L. Zhang, Y. Xie, Y. Zheng, W. Xue, X. Zheng, X. Xu, "The challenges and countermeasures of blockchain in finance and economics," Syst. Res. Behav. Sci., 37(4): 691-698, 2020.

[3] M. U. CHELLADURAI, S. Pandian, K. Ramasamy, "A blockchain based patient centric EHR storage and integrity management for e-Health systems," Heal. Policy Technol., 10(4): 100513, 2021.

[4] S. Shamshad, K. Mahmood, S. Kumari, C. M. Chen, "A secure blockchain-based e-health records storage and sharing scheme," J. Inf. Secur. Appl., 55: 102590, 2020.

[5] E. Bandara, D. Tosh, P. Foytik, S. Shetty, N. Ranasinghe, K. De Zoysa, "Tikiri—Towards a lightweight blockchain for IoT," Futur. Gener. Comput. Syst., 119: 154-165, 2021.

[6] U. Majeed, L. U. Khan, I. Yaqoob, S. M. A. Kazmi, K. Salah, C. S. Hong, "Blockchain for IoT-based smart cities: Recent advances, requirements, and future challenges," J. Netw. Comput. Appl., 181: 103007, 2021.

[7] P. Asghari, A. M. Rahmani, H. H. S. Javadi, "Internet of Things applications: A systematic review," Comput. Networks, 148: 241-261, 2019.

[8] P. Centobelli, R. Cerchione, P. Del Vecchio, E. Oropallo, G. Secundo, "Blockchain technology for bridging trust, traceability and transparency in circular supply chain," Inf. Manag., 59(7): 103508, 2021.

[9] B. Wang, W. Luo, A. Zhang, Z. Tian, Z. Li, "Blockchain-enabled circular supply chain management: A system architecture for fast fashion," Comput. Ind., 123: 103324, 2020.

[10] F. Luo, Z. Y. Dong, G. Liang, J. Murata, Z. Xu, "A distributed electricity trading system in active distribution networks based on multi-agent coalition and blockchain," IEEE Trans. Power Syst., 34(5): 4097-4108, 2018.

[11] J. Wang, Q. Wang, N. Zhou, Y. Chi, "A novel electricity transaction mode of microgrids based on blockchain and continuous double auction," Energies, 10(12): 1971, 2017.

[12] M. Raikwar, S. Mazumdar, S. Ruj, S. Sen Gupta, A. Chattopadhyay, and K. Y. Lam, "A blockchain framework for insurance processes,"

in Proc. 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS): 1-4, 2018.

[13] S. Gupta, A. Gupta, I. Y. Pandya, A. Bhatt, K. Mehta, "End to end secure e-voting using blockchain & quantum key distribution," Mater. Today Proc., 80: 3363-3370, 2023.

[14] X. Yang, X. Yi, S. Nepal, A. Kelarev, F. Han, "Blockchain voting: Publicly verifiable online voting protocol without trusted tallying authorities," Futur. Gener. Comput. Syst., 112: 859-874, 2020.

[15] S. N. Mohanty et al., "An efficient Lightweight integrated Blockchain (ELIB) model for IoT security and privacy," Futur. Gener. Comput. Syst., 102: 1027-1037, 2020.

[16] Y. L. Gao, X. B. Chen, Y. L. Chen, Y. Sun, X. X. Niu, Y. X. Yang, "A secure cryptocurrency scheme based on post-quantum blockchain," IEEE Access, 6: 27205–27213, 2018.

[17] Y. Liu, G. Xu, "Fixed degree of decentralization DPoS consensus mechanism in blockchain based on adjacency vote and the average fuzziness of vague value," Comput. Networks, 199: 108432, 2021.

[18] N. Alzahrani, N. Bulusu, "Towards true decentralization: A blockchain consensus protocol based on game theory and randomness," in Proc. International Conference on Decision and Game Theory for Security: 465–485, 2018.

[19] E. Bandara, X. Liang, P. Foytik, S. Shetty, N. Ranasinghe, K. De Zoysa, "Rahasak—Scalable blockchain architecture for enterprise applications," J. Syst. Archit., 116: 102061, 2021.

[20] A. Dorri, S. S. Kanhere, R. Jurdak, P. Gauravaram, "LSB: A Lightweight Scalable Blockchain for IoT security and anonymity," J. Parallel Distrib. Comput., 134: 180-197, 2019.

[21] M. Al-Bassam, A. Sonnino, S. Bano, D. Hrycyszyn, G. Danezis, "Chainspace: A sharded smart contracts platform," arXiv Prepr. arXiv1708.03778, 2017.

[22] A. I. Sanka, M. H. Chowdhury, R. C. C. Cheung, "Efficient high-performance FPGA-Redis hybrid NoSQL caching system for blockchain scalability," Comput. Commun., 169: 81–91, 2021.

[23] M. Muzammal, Q. Qu, B. Nasrulin, "Renovating blockchain with distributed databases: An open source system," Futur. Gener. Comput. Syst., 90: 105-117, 2019.

[24] Q. Qu, I. Nurgaliev, M. Muzammal, C. S. Jensen, J. Fan, "On spatio-temporal blockchain query processing," Futur. Gener. Comput. Syst., 98: 208-218, 2019.

[25] S. Linoy, H. Mahdikhani, S. Ray, R. Lu, N. Stakhanova, A. Ghorbani, "Scalable privacy-preserving query processing over ethereum blockchain," in Proc. 2nd IEEE Int. Conf. Blockchain, Blockchain 2019: 398–404, 2019.

[26] L. Zeng, W. Qiu, X. Wang, H. Wang, Y. Yao, Z. Yu, "Transaction-based Static Indexing Method to Improve the Efficiency of Query on the Blockchain," in Proc. 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA): 780–784, 2021.

[27] C. Riegger, T. Vinçon, I. Petrov, "Efficient data and indexing structure for blockchains in enterprise systems," in Proc. the 20th International Conference on Information Integration and Web-based Applications & Services: 173-182, 2018.

[28] V. Buterin, "Ethereum Whitepaper," 2013.

[29] J. Rubin, M. Naik, N. Subramanian, "Merkelized abstract syntax trees," XP055624837, Dec, 16(3), 2014.

[30] E. Lombrozo, J. Lau, P. Wuille, "Segregated witness (consensus layer)," Bitcoin Core Dev. Team, Tech. Rep. BIP, 141, 2015.

[31] Cash B. "Bitcoin Cash," 2017.

[32] M. A. Javarone, C. S. Wright, "From Bitcoin to Bitcoin Cash: a network analysis," in Proc. the 1st Workshop on Cryptocurrencies and Blockchains for Distributed Systems: 77-81, 2018.

[33] W. K. Chan, J. J. Chin, V. T. Goh, "Simple and scalable blockchain with privacy," J. Inf. Secur. Appl., 58: 102700, 2021.

[34] Z. Xu, S. Han, L. Chen, "CUB, a consensus unit-based storage scheme for blockchain system," in Proc. IEEE 34th International Conference on Data Engineering (ICDE): 173-184, 2018.

[35] X. Dai, J. Xiao, W. Yang, C. Wang, H. Jin, "Jidar: A jigsaw-like data reduction approach without trust assumptions for bitcoin system," in Proc. 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS): 1317-1326, 2019.

[36] Y. Xu, Y. Huang, "Segment blockchain: A size reduced storage mechanism for blockchain," IEEE Access, 8: 17434–17441, 2020.

[37] D. Y. Jia, J. C. Xin, Z. Q. Wang, H. Lei, G. R. Wang, "SE-Chain: A scalable storage and efficient retrieval model for blockchain," J. Comput. Sci. Technol., 36(3): 693-706, 2021.

[38] D. Ding, X. Jiang, J. Wang, H. Wang, X. Zhang, Y. Sun, "Txilm: Lossy block compression with salted short hashing," arXiv Prepr. arXiv1906.06500, 2019.

[39] J. Poon, T. Dryja, "The bitcoin lightning network: Scalable off-chain instant payments," 2016.

[40] Network-Fast R. "Cheap, Scalable Token Transfers for Ethereum," 2018.

[41] "µRaiden. A Payment Channel Framework for Fast and Free Off-Chain ERC20 Token Transfers," 2018.

[42] "Trinity. Universal Off-Chain Scaling Solution for Neo," 2018.

[43] C. Lin, N. Ma, X. Wang, J. Chen, "Rapido: Scaling blockchain with multi-path payment channels," Neurocomputing, 406: 322-32, 2020.

[44] J. Poon, V. Buterin, "Plasma: Scalable autonomous smart contracts," White Pap., 1-47, 2017.

[45] M. Mallaki, B. Majidi, A. Peyvandi, A. Movaghar, "Off-chain management and state-tracking of smart programs on blockchain for secure and efficient decentralized computation," Int. J. Comput. Appl., 44(9): 822–829, 2022.

[46] R. van der Meyden, "On the specification and verification of atomic swap smart contracts," in Proc. 2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC): 176-179, 2019.

[47] Q. Hu, B. Yan, Y. Han, J. Yu, "An improved delegated proof of stake consensus algorithm," Procedia Comput. Sci., 187: 341-346, 2021.

[48] D. Larimer, "Delegated Proof of Stake (DPoS), Bitshare Whitepaper," 2014.

[49] M. Castro, B. Liskov, "Practical byzantine fault tolerance," in OSDI, 99(1999): 173-186,1999.

[50] M. Lokhava et al., "Fast and secure global payments with Stellar," in Proc. the 27th ACM Symposium on Operating Systems Principles: 80-96, 2019.

[51] M. Bareis, M. Di Angelo, G. Salzer, "Functional differences of neo and ethereum as smart contract platforms," in Proc. International Congress on Blockchain and Applications: 13-23, 2020.

[52] H. Da and E. Zhang, "Neo cryptocurrency," 2018.

[53] S. De Angelis, L. Aniello, R. Baldoni, F. Lombardi, A. Margheri, V. Sassone, "PBFT vs proof-of-authority: Applying the CAP theorem to permissioned blockchain," 2018.

[54] "Proof of Authority Whitepaper," 2018.

[55] Y. Gilad, R. Hemo, S. Micali, G. Vlachos, N. Zeldovich, "Algorand: Scaling byzantine agreements for cryptocurrencies," in Proc. the 26th symposium on operating systems principles: 51-68, 2017.

[56] M. Divya, N. B. Biradar, "IOTA-next generation block chain," Int. J. Eng. Comput. Sci, 7(04): 23823-23826, 2018.

[57] C. LeMahieu, "Nano: A feeless distributed cryptocurrency network," 16: 17, 2018.

[58] T. Zhou, X. Li, H. Zhao, "DLattice: A permission-less blockchain based on DPoS-BA-DAG consensus for data tokenization," IEEE Access, 7: 39273–39287, 2019.

[59] L. Baird, "The swirlds hashgraph consensus algorithm: Fair, fast, byzantine fault tolerance," Swirlds Tech. Reports SWIRLDS-TR-2016-01, Tech. Rep, 34: 9-11, 2016.

[60] J. He, G. Wang, G. Zhang, J. Zhang, "Consensus mechanism design based on structured directed acyclic graphs," Blockchain Res. Appl., 2(1): 100011, 2021.

[61] G. Wang, J. Zhang, G. Zhang, J. He, "Consensus mechanism design based on structured directed acyclic graphs," arXiv, 2019.

[62] P. Otte, M. de Vos, J. Pouwelse, "TrustChain: A Sybil-resistant scalable blockchain," Futur. Gener. Comput. Syst., 107: 770-780, 2020.

[63] L. Luu, V. Narayanan, C. Zheng, K. Baweja, S. Gilbert, P. Saxena, "A secure sharding protocol for open blockchains," in Proc. the 2016 ACM SIGSAC Conference on Computer and Communications Security: 17–30, 2016.

[64] E. Kokoris-Kogias, P. Jovanovic, L. Gasser, N. Gailly, E. Syta, B. Ford, "Omniledger: A secure, scale-out, decentralized ledger via sharding," in Proc. 2018 IEEE Symposium on Security and Privacy (SP): 583–598, 2018.

[65] M. Zamani, M. Movahedi, M. Raykova, "Rapidchain: Scaling blockchain via full sharding," in Proc. 2018 ACM SIGSAC Conference on Computer and Communications Security: 931–948, 2018.

[66] P. Barrett, "Technical Whitepaper," Zilliqa, 1-8, 2017.

[67] H. Team "Harmony, Technical Whitepaper," 2018.

[68] J. Wang, H. Wang, "Monoxide: Scale out blockchain with asynchronous consensus zones," in Proc. 16th USENIX Symp. Networked Syst. Des. Implementation, NSDI 2019: 95-112, 2019.

[69] Y. Liu, J. Liu, D. Li, H. Yu, Q. Wu, "Fleetchain: A secure scalable and responsive blockchain achieving optimal sharding," in Proc. International Conference on Algorithms and Architectures for Parallel Processing: 409-425, 2020.

[70] Z. Hong, S. Guo, P. Li, W. Chen, "Pyramid: A layered sharding blockchain system," in Proc. IEEE INFOCOM 2021-IEEE Conference on Computer Communications: 1-10, 2021.

[71] C. Huang et al., "RepChain: A reputation-based secure, fast, and high incentive blockchain system via sharding," IEEE Internet Things J., 8(6): 4291-4304, 2020.

[72] H. Dang, T. T. A. Dinh, D. Loghin, E. C. Chang, Q. Lin, B. C. Ooi, "Towards scaling blockchain systems via sharding," in Proc. 2019 international conference on management of data: 123–140, 2019.

[73] C. Fan, S. Ghaemi, H. Khazaei, P. Musilek, "Performance evaluation of blockchain systems: A systematic survey," IEEE Access, 8: 126927–126950, 2020.

[74] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Decentralized Bus. Rev., 2008.

[75] Y. C. Liang, "Blockchain for dynamic spectrum management," Dyn. Spectr. Manag. From Cogn. Radio to Blockchain Artif. Intell., 121–146, 2020.

[76] S. Popov, "The tangle," White Pap., 1(3): 30, 2018.

[77] R. G. Brown, J. Carlyle, I. Grigg, M. Hearn, "Corda: an introduction," R3 CEV, August, 1(15): 14, 2016.

[78] "Radix DeFi White Paper," 1–31, 2020.

[79] "The Ethereum project, the modified merkle patricia tree," 2020.

[80] O. Sury, R. Edmonds, "Edwards-Curve Digital Security Algorithm (EdDSA) for DNSSEC," RFC 8080 (Proposed Standard)., 2017.

[81] G. Maxwell, A. Poelstra, "Borromean ring signatures," Accessed Jun, 8: 2019, 2015.

[82] L. Lamport, "Constructing digital signatures from a one-way function," 1979.

[83] J. Reed, "Litecoin: An introduction to litecoin cryptocurrency and litecoin mining," CreateSpace Independent Publishing Platform, 2017.

[84] D. Schwartz, N. Youngs, A. Britto, "The ripple protocol consensus algorithm," Ripple Labs Inc. White Pap., 5(8): 151, 2014.

[85] "EOSIO, An Open-source Blockchain Platfom,"2018.

[86] D. Khovratovich, J. Law, "Sovrin: digital identities in the blockchain era," Github Commit by jasonalaw Oct., 17: 38-99, 2017.

[87] "LTO Network,Blockchain for Decentralized Workflows," 2019.

[88] "Holochain: Scalable Agent-Centric Distributed Computing," 2018.

[89] "Monet Network,"2018.

[90] E. Androulaki et al., "Hyperledger fabric: a distributed operating system for permissioned blockchains," in Proc. the thirteenth EuroSys conference: 1-15, 2018.

[91] S. Ghimire, H. Selvaraj, "A survey on bitcoin cryptocurrency and its mining," in Proc. 2018 26th International Conference on Systems Engineering (ICSEng): 1-6, 2018.

[92] D. Ongaro, J. Ousterhout, "In search of an understandable consensus algorithm," in Proc. 2014 {USENIX} Annual Technical Conference ({USENIX}{ATC} 14): 305-319, 2014.

[93] C. Dannen, "Introducing Ethereum and solidity",Berkeley, 1: 159-160, 2017.

[94] A. Hafid, A. S. Hafid, M. Samih, "Scaling blockchains: A comprehensive survey," IEEE Access, 8: 125244–125262, 2020.

[95] Q. Zhou, H. Huang, Z. Zheng, J. Bian, "Solutions to scalability of blockchain: A survey," IEEE Access, 8: 16440–16455, 2020.

[96] M. H. Nasir, J. Arshad, M. M. Khan, M. Fatima, K. Salah, R. Jayaraman, "Scalable blockchains—A systematic review," Futur. Gener. Comput. Syst., 126: 136-162, 2022.

[97] G. Yu, X. Wang, K. Yu, W. Ni, J. A. Zhang, R. P. Liu, "Survey: Sharding in blockchains," IEEE Access, 8: 14155–14181, 2020.

[98] Q. Wang, J. Yu, S. Chen, Y. Xiang, "SoK: Diving into DAG-based blockchain systems," arXiv Prepr. arXiv2012.06128, 2020.

[99] D. P. Oyinloye, J. Sen Teh, N. Jamil, M. Alawida, "Blockchain consensus: An overview of alternative protocols," Symmetry (Basel)., 13(8): 1363, 2021.

[100] J. Garzik, "Block size increase to 2MB," Bitcoin Improv. Propos., 102, 2015.

[101] B. Yu, X. Li, H. Zhao, "PoW-BC: A PoW consensus protocol based on block compression.," KSII Trans. Internet Inf. Syst., 15(4), 2021.

[102] U. Nadiya, K. Mutijarsa, C. Y. Rizqi, "Block summarization and compression in bitcoin blockchain," in Proc. 2018 International Symposium on Electronics and Smart Devices (ISESD): 1-4, 2018.

[103] S. Kim, Y. Kwon, S. Cho, "A survey of scalability solutions on blockchain," in Proc. 2018 International Conference on Information and Communication Technology Convergence (ICTC): 1204-1207, 2018.

[104] C. Decker, R. Wattenhofer, "Bitcoin transaction malleability and MtGox," in Proc. European Symposium on Research in Computer Security: 313–326, 2014.

[105] A. Back et al., "Enabling blockchain innovations with pegged sidechains," 72: 201-24 , 2014.

[106] R. A. N. Yaxuan, N. I. U. Yixin, C. Siyun, "More is less: Why multiple payment mechanism impairs individual donation," Acta Psychol. Sin., 53(4): 413, 2021.

[107] S. Bistarelli, C. Pannacci, F. Santini, "CapBAC in Hyperledger Sawtooth," in Proc. IFIP International Conference on Distributed Applications and Interoperable Systems: 152-169, 2019.

[108] "Proof of Capacity,"2018.

[109] Y. Lai, "NEM White paper," Imid 2009, (159679): 1069-1072, 2018.

[110] K. Karantias, A. Kiayias, D. Zindros, "Proof-of-burn," in Proc.

International Conference on Financial Cryptography and Data Security: 523–540, 2020.

[111] Q. Wang et al., "Security analysis on dBFT protocol of NEO," in Proc. International Conference on Financial Cryptography and Data Security: 20–31, 2020.

[112] "Steemit, A Blockchain-Based Blogging and Social Media project,"2017.

[113] "Ark Blockchain Framework,"2019.

[114] "Cardano, A Blockchain project," 2017.

[115] "Lisk Blockchain Application Platform,"2017.

[116] S. Tang, Z. Wang, J. Jiang, S. Ge, G. Tan, "Improved PBFT algorithm for high-frequency trading scenarios of alliance blockchain," Sci. Rep., 12(1): 4426, 2022.

[117] E. Buchman, "Tendermint: Byzantine fault tolerance in the age of blockchains," 2016.

[118] "Ontology, A blockchain for self-sovereign ID and DATA," 2018.

[119] P. Khahulin, I. Barinov, V. Baranov, "POA network whitepaper; technical report," 2018.

[120] "VeChainThor public Blockchain", 2019.

[121] A. Churyumov, "Byteball: A decentralized system for storage and transfer of value," 2016.

[122] Y. Sompolinsky, Y. Lewenberg, A. Zohar, "SPECTRE: a fast and scalable cryptocurrency protocol," IACR Cryptol. ePrint Arch., 2016.

[123] X. Boyen, C. Carr, T. Haines, "Graphchain: A blockchain-free scalable decentralised ledger," in Proc. the 2nd ACM Workshop on Blockchains, Cryptocurrencies, and Contracts: 21-33, 2018.

[124] G. Srivastava, A. D. Dwivedi, R. Singh, "PHANTOM protocol as the new crypto-democracy," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics): 499–509, 2018.

[125] H. Gupta, D. Janakiram, "Cdag: A serialized blockdag for permissioned blockchain," arXiv Prepr. arXiv1910.08547, 2019.

[126] C. Li, P. Li, D. Zhou, W. Xu, F. Long, A. Yao, "Scaling nakamoto consensus to thousands of transactions per second," arXiv Prepr. arXiv1805.03870, 2018.

[127] T. Y. Chen, W. N. Huang, P. C. Kuo, H. Chung, T. W. Chao, "DEXON: a highly scalable, decentralized DAG-based consensus algorithm," arXiv Prepr. arXiv1811.07525, 2018.

[128] X. Fu, H. Wang, P. Shi, X. Zhang, "Teegraph: A Blockchain consensus algorithm based on TEE and DAG for data sharing in IoT," J. Syst. Archit., 122: 102344, 2022.

[129] "Red-Black Merkle Tree,"2015.

[130] A. Manuskin, M. Mirkin, I. Eyal, "Ostraka: Secure blockchain scaling by node sharding," in Proc. 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW): 397-406, 2020.

[131] E. K. Kogias, P. Jovanovic, N. Gailly, I. Khoffi, L. Gasser, B. Ford, "Enhancing bitcoin security and performance with strong consistency via collective signing," in Proc. 25th {usenix} security symposium ({usenix} security 16): 279-296, 2016.

[132] S. Sen, M. J. Freedman, "Commensal cuckoo: Secure group partitioning for large-scale services," ACM SIGOPS Oper. Syst. Rev., 46(1): 33-39, 2012.

[133] I. Eyal, A. E. Gencer, E. G. Sirer, R. Van Renesse, "Bitcoin-ng: A scalable blockchain protocol," in Proc. 13th {USENIX} symposium on networked systems design and implementation ({NSDI} 16): 45-59, 2016.

[134] B. S. Reddy, "securePrune: Secure block pruning in UTXO based blockchains using Accumulators," in Proc. 2021 International Conference on COMmunication Systems & NETworkS (COMSNETS): 174–8, 2021.

[135] V. Bagaria, S. Kannan, D. Tse, G. Fanti, P. Viswanath, "Prism: Deconstructing the blockchain to approach physical limits," in Proc. the 2019 ACM SIGSAC Conference on Computer and Communications Security: 585-602, 2019.

[136] H. Yu, I. Nikolić, R. Hou, P. Saxena, "Ohie: Blockchain scaling made simple," in Proc. 2020 IEEE Symposium on Security and Privacy (SP): 90-105, 2020.

[137] Z. Yin et al., "Streamnet: A dag system with streaming graph computing," in Proc. the Future Technologies Conference: 499-522, 2020.

[138] J. Niu, "Eunomia: A permissionless parallel chain protocol based on logical clock," arXiv Prepr. arXiv1908.07567, 2019.

[139] M. J. Amiri, D. Agrawal, A. El Abbadi, "Caper: a cross-application permissioned blockchain," Proc. VLDB Endow., 12(11): 1385-1398, 2019.

[140] A. Gągol, D. Leśniak, D. Straszak, M. Świętek, "Aleph: Efficient atomic broadcast in asynchronous networks with byzantine nodes," in Proc. the 1st ACM Conference on Advances in Financial Technologies: 214–228, 2019.

[141] L. N. Nguyen, T. D. T. Nguyen, T. N. Dinh, M. T. Thai, "Optchain: optimal transactions placement for scalable blockchain sharding," in Proc. 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS): 525-535, 2019.

[142] M. J. Amiri, D. Agrawal, A. El Abbadi, "Sharper: Sharding permissioned blockchains over network clusters," in Proc. the 2021 International Conference on Management of Data: 76-88, 2021.

[143] H. Chen, Y. Wang, "Sschain: A full sharding protocol for public blockchain without data migration overhead," Pervasive Mob. Comput., 59: 101055, 2019.

Biographies

Alemeh Matani received the B.Sc. degree from the University of Mazandaran, Iran, in 2013, and the M.Sc. degree from Kerman Graduate University of Technology, Iran, in 2015, both in Information Technology Engineering. She is currently pursuing the Ph.D. degree in Computer Engineering at the Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran. Her current research interests include blockchain, the Internet of Things (IoT) and distributed computing systems.

- Email: alemehmatani@gmail.com
- ORCID: 0009-0009-3259-736X
- Web of Science Researcher ID: JFQ-6782-2023
- Scopus Author ID: 57212064760
- Homepage: NA

Amir Sahafi received the B.Sc. degree from Shahed University of Tehran, Iran in 2005, M.Sc. and Ph.D. degrees both from Science and Research Branch of Islamic Azad University, Tehran, Iran, in 2007 and 2012, all in Computer Engineering. He is an Assistant Professor in Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran. His current research interests are Distributed and Cloud computing.

- Email: sahafi@iau.ac.ir
- ORCID: 0000-0002-6555-670X
- Web of Science Researcher ID: AAS-1208-2021
- Scopus Author ID: 24528878600
- Homepage: https://stb.iau.ir/faculty/a-sahafi

Ali Broumandnia was born in Isfahan, Iran. He received the B.Sc. degree from Isfahan university of Technology 1991, M.Sc. degree from Iran University of Science and Technology in 1995, both in Hardware Engineering and Ph.D. degree of Computer Engineering from Tehran Islamic Azad University-Science and Research Branch in 2006. From 1993 through 1995, he worked on intelligent transportation control with image processing and designed the Automatic License Plate Recognition for Tehran Control Traffic Company. He has published over 30 computer books, journal and conference papers. He is interested in Persian/Arabic character recognition and segmentation, Persian/Arabic document segmentation, medical imaging, signal and image processing, and wavelet analysis. He is reviewer of some International journals and conferences.

- Email: broumandnia@azad.ac.ir
- ORCID: 0000-0001-5145-2013
- Web of Science Researcher ID: I-6383-2018
- Scopus Author ID: 23003455800
- Homepage: https://stb.iau.ir/faculty/a-broumandnia

How to cite this paper:
A. Matani, A. Sahafi, A. Broumandnia, "A comprehensive review on blockchain scalability," J. Electr. Comput. Eng. Innovations, 12(1): 187-216, 2024.

DOI: 10.22061/jecei.2023.9975.670

URL: https://jecei.sru.ac.ir/article_2000.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Optimum Spectral Indices for Water Bodies Recognition Based on Genetic Algorithm and Sentinel-2 Satellite Images

H. Karim Tabahfar, F. Tabib Mahmoudi *

Department of Geomatics, Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:
Received 07 August 2023
Reviewed 17 September 2023
Revised 24 October 2023
Accepted 05 November 2023

Keywords:
Genetic algorithm
Spectral indices
Water bodies
Classifier
Optimization

*Corresponding Author's Email Address:
fmahmooudi@sru.ac.ir

Abstract

Background and Objectives: Considering the drought and global warming, it is very important to monitor changes in water bodies for surface water management and preserve water resources in the natural ecosystem. For this purpose, using the appropriate spectral indices has high capabilities to distinguish surface water bodies from other land covers. This research has a special consideration to the effect of different types of land covers around water bodies. For this reason, two different water bodies, lake and wetland, have been used to evaluate the implementation results.

Methods: The main objective of this research is to evaluate the capabilities of the genetic algorithm in optimum selection of the spectral indices extracted from Sentinel-2 satellite image due to distinguish surface water bodies in two case studies: 1) the pure water behind the Karkheh dam and 2) the Shadegan wetland having water mixed with vegetation. In this regard, the set of optimal indices is obtained with the genetic algorithm followed by the support vector machine (SVM) classifier.

Results: The evaluation of the classification results based on the optimum selected spectral indices showed that the overall accuracy and Kappa coefficient of the recognized surface water bodies are 98.18 and 0.9827 in the Karkheh dam and 98.04 and 0.93 in Shadegan wetland, respectively. Evaluation of each of the spectral indices measured in both study areas was carried out using quantitative decision tree (DT) classifier. The best obtained DT classification results show the improvements in overall accuracy by 1.42% in the Karkheh Dam area and 1.56% in the Shadegan Wetland area based on the optimum selected indices by genetic algorithm followed by SVM classifier. Moreover, the obtained classification results are superior compared with Random Forest classifier using the optimized set of spectral features.

Conclusion: Applying the genetic algorithm on the spectral indices was able to obtain two optimal sets of effective indices that have the highest amount of accuracy in classifying water bodies from other land cover objects in the study areas. Considering the collective performance, genetic algorithm selects an optimal set of indices that can detect water bodies more accurately than any single index.

Introduction

Surface water bodies those are usually the sources of fresh water, have very important role in both human life and environmental protection [1]. The water supplies help to maintain biodiversity in water and wetland ecosystems. This is not only vital for ecosystems as a key component of the hydrological cycle, but also related to every aspect of human life such as drinking water,

agriculture, electricity generation, transportation and industrial purposes. Changes in the characteristics of surface water bodies may lead to the occurrence of severe disasters such as floods, droughts and even the spread of water-related diseases, all of which have consequences for the safety of human life and property, [2], [3].

In order to reach the spatial distribution and expansion of geographic information, it is necessary to accurately recognize the surface water bodies based on satellite remote sensing data [4]-[8]. One of the most widely used remote sensing data analysis for detecting surface water bodies are the methods based on using spectral indices that identify water bodies based on water absorption bands in different wavelengths of the electromagnetic spectrum [9]-[16]. The results of surface water bodies recognition based on remote sensing data are important in various scientific fields, including research in the evaluation of existing and future climate models, agriculture, river dynamics, wetland studies, watershed analysis, flood mapping and environmental monitoring.

In a research conducted by Emami et al, four different water indices including the WRI (water ratio index), AWEI (automatic water extraction index), NDWI (normalized difference water index) and NDVI (normalized difference vegetation index) are used to reveal and evaluate the spatial-temporal changes in the water level of Uremia lake in the time period 2002 to 2016 and based on the Landsat satellite images [16]. The water ratio index is defined according to the spectral reflectance of water in the green and red bands in comparison with near-infrared and middle infrared regions. Values greater than one of this index are considered as water pixels. The automatic water extraction index is one of the indices that are mostly used for the extraction of water bodies in urban areas. This index is used to remove dark pixels and identify water surfaces with high accuracy in urban and mountain areas where the shadow problem prevents correct identification [16]. The normalized difference water index is sensitive to water changes. This index is calculated using near-infrared and short-wavelength infrared reflections. The range of changes of this index is between -1 and +1 and water has positive values [9], [17], [18]. The normalized difference vegetation index is defined by the amount of reflective energy in the red and near-infrared bands and has values between -1 and +1 [19]. In this index, water, snow and ice have negative values. One of the major errors affecting the values of this index is clouds and atmospheric pollution such as smoke, fog and dust, which increase or decrease the values of this index. Obviously, the absorption of infrared rays by water and its intense reflection by vegetation and soil provides an ideal combination of these bands for extracting water bodies. In 2018, Haung et al. investigated some water-related indices in the Poyang Lake region of China [20]. In this study, while examining different types of water indices, it was concluded that the improved normalized difference water index MNDWI is more reliable than the NDWI. Because, this index uses the SWIR spectral band and is less sensitive than NIR band to the density of sediments and other optically active components in water [20].

Thresholding is one of the most important methods in using spectral indices for water extraction. Based on the reflectance characteristics of water, the values of NDWI and MNDWI for water are usually greater than zero. Therefore, zero is often used to as threshold for extracting water from spectral index images. However, by fine-tuning the threshold values, better extraction results can usually be achieved [21].

In 2013, Fisher et al. investigated water indices extracted from SPOT-5 satellite images over the New South Wales region of Australia. These indices included three different types of normalized differences water index named $NDWI_{gao}$, $NDWI_{xu}$ and $NDWI_{McFeeters}$. In this research, by calculating the normalized differences between two bands and then applying a threshold, several different water indices are produced and their performance for the classification of water bodies in remote sensing images have been investigated. The $NDWI_{McFeeters}$ index consists of the combination of green and near-infrared bands, the $NDWI_{gao}$ index consists of the combination of near-infrared and short-wave infrared, and the $NDWI_{xu}$ index consists of the combination of green and short-wave infrared bands. The results of investigating the performance of these three indices in a coastal image containing abundant water areas showed that the $NDWI_{McFeeters}$ separates water and non-water effects well and has a strong negative value in vegetation, while the values of this index in water areas is greater than zero. Water values are higher in the $NDWI_{xu}$, although the vegetation response also has large values. The $NDWI_{gao}$ provides a very poor separation of water and non-water, and both vegetation and water have large positives [22].

Meta-heuristic methods such as genetic algorithm or SWARM intelligence algorithms are also successfully used in some research for optimum selection of features or image bands [23]-[25]. In the research conducted in this paper, instead of comparing the limited amount of spectral water indices, genetic algorithm is used for optimum selection from a large set of spectral indices. The main contributions of this research are as following:

- Evaluating the capabilities of genetic algorithm for optimum selection of water indices.
- Considering the impact of various land covers (especially vegetation) in water body recognition by

using two types of study areas with different characteristics.

- Discussing the relationships between the type of features selected by the genetic algorithm and the difference in land covers around each of the study areas.
- Comparing the efficiency results of each individual water index and optimum set of indices selected by genetic algorithm in water body recognition.

Materials and Method

One of the important issues related to the surface water bodies detection is the use of optimal indices in the detection and separation of water from other land covers. Karkheh dam with geographical location (21°29′32″ N, 36°07′48″ E) and Shadegan wetland with geographical location (58°38′30″ N, 52°39′48″E) in Khuzestan province are two samples of surface water bodies those are considered as the study areas of this research. Karkheh dam is one of the largest dams in the world and is the largest dam in the Middle East, which was built on the Karkheh River in Andimshek city of Khuzestan province. The Karkheh River is the third largest river in Iran after the Karun and Dez rivers from the water supply point of view.

Shadegan international wetland is one of the large wetlands, which is located in the southwest of Iran and in the south of Shadegan city in Khuzestan province. The dams construction, not enough water supply of the wetland, the discharge of polluted effluents such as sugarcane fields, fish farming in Khuzestan, the passage of oil pipelines and the activity of 30 petrochemical units have included this unique wetland in the red list of international wetlands since 1993.

The water of this international wetland is supplied from the rivers of Jarahi and Karun, as well as the tides of the Persian Gulf, which, despite the fact that its fresh water part is seasonal; the salt water of this wetland is permanent.

In this research, images those are taken by the medium spatial resolution Sentinel-2 satellite sensor have been used. The capabilities of this satellite sensor are the multi-spectral imaging using 13 spectral bands in the visible and infrared ranges. This satellite has regular global coverage and also covers the coastal waters and the entire Mediterranean Sea.

Two Sentinel-2 satellite images from the study areas of Karkheh Dam and Shadegan wetland on December 2020 were used in this study.

An attempt was made to use images with the right weather conditions, without dust or cloud covers. Sentinel-2 images from the study areas of this research are shown in Fig. 1.

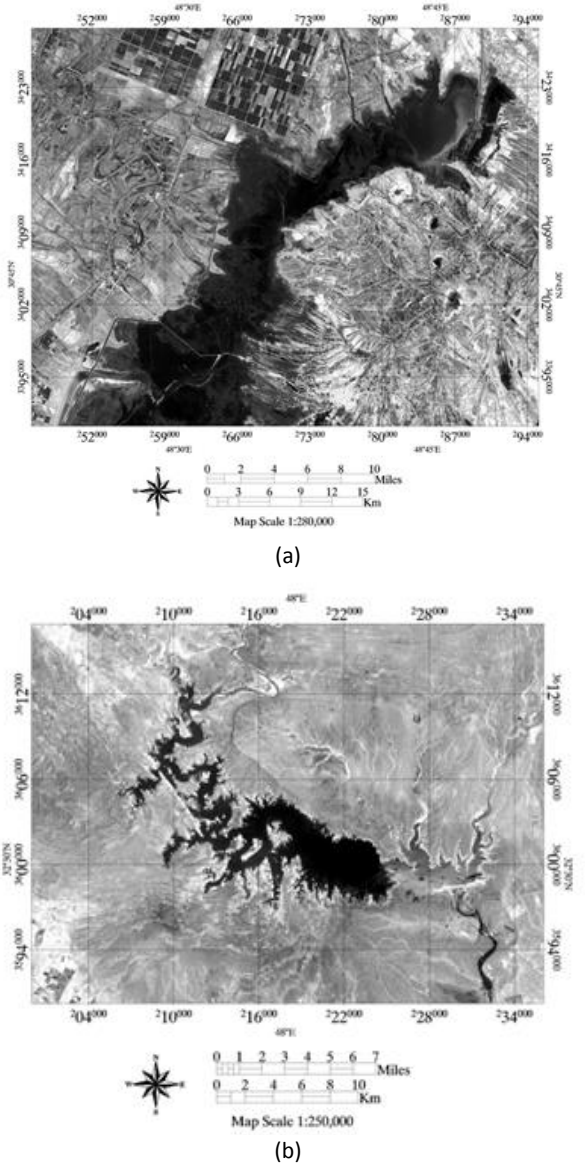


Fig. 1: Sentinel-2 satellite image of a) Shadegan wetland, b) Karkheh Dam.

As shown in Fig. 2, the proposed method in this research in the first stage consists of performing pre-processing (atmospheric correction, normalization and cropping) on Sentinel-2 satellite images obtained from the study areas.

In the second stage, after measuring the large set of spectral indices related to the surface water bodies, the genetic algorithm followed by the support vector machine classifier is applied for determining the optimal set of spectral indices for surface water detection. In parallel, the decision tree classifier is performed in order to evaluate the capability of each spectral index and comparing it with the optimization results.

A. Spectral Water Indices

According to the research background, the names and definitions of the spectral indices used in the detection and classification of surface water bodies from remote sensing images are mentioned in this section and the

mathematical formulas of these indices are also presented in Table 1 [26], [27].

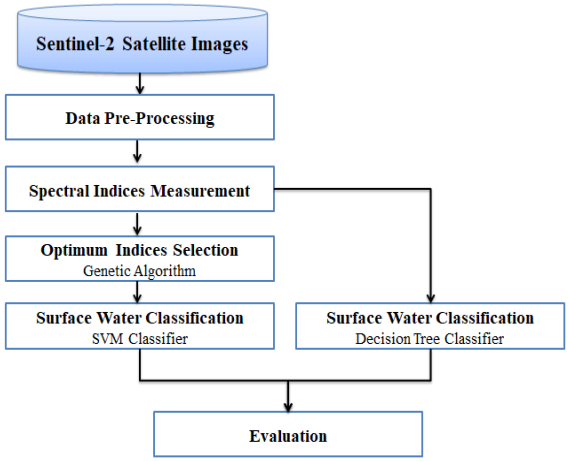


Fig. 2: Proposed method for surface water bodies classification.

- NDWI is the normalized difference water index for distinguishing water areas with little vegetation.
- MNDWI is the modified normalized difference water index.
- AWEIsh is automatic water extraction index shadow which is used for water extraction in urban areas.
- AWEInsh is automatic water extraction index no-shadow which is used more often in urban areas with fewer shadows.
- WRI is the water ratio index which is the ratio between total reflection of red and green bands to middle and near infrared bands.
- NDVI is the normalized difference vegetation index as the most widely used vegetation index, which is also used to identify water areas.
- NDMI is the normalized difference moisture index which is used to determine the soil moisture.
- NDPI is the normalized difference index of water ponds which is used to identify water areas in wetlands.
- TCW is wet index used to determine pixels with high humidity.
- WI2015 is the developed water index, which is one of the useful indices in determining water covers due to its high accuracy.
- NDSI is the normalized difference snow index which is used to identify snow covers from other water bodies.
- NDTI is the normalize difference turbidity index which is estimated using the spectral reflectance values of the water pixels to estimate the turbidity in water bodies.
- SWI is simple water index which expresses the simple relationship between blue and mid-infrared band reflections.

Table 1: Name and mathematical formulas of the utilized spectral indices in the paper

Index	Mathematical Formula	Index	Mathematical Formula
NDWI	(Green-NIR)/(Green+NIR)	TCW	0.1511NIR × BLUE - 0.7117 + 0.1973 × SWIR1 × GREEN - 0.4559 + 0.3283 × SWIR2 × RED
MNDWI	(Green-MIR)/(Green+MIR)	TCW2	0.0315×Bblue+0.1973×Bgreen+0.3279×Bred+0.3406×Bnir-0.7112×Bswir1-0.4572×Bswir2
AWEIsh	Blue+2.5×Green-1.5×(NIR+SWIR1)-0.25× SWIR2	NDWI (Blue)	(Blue-NIR)/(Blue+NIR)
AWEInsh	4×(Green-SWIR1)-(0.25×NIR+2.75×SWIR2)	NDWI (Red)	(Red-NIR)/(Red+NIR)
WRI	(Green+Red)/(NIR+SWIR1)	WI2015	1.7204 + 171(GREEN) + 3(RED) + 70(NIR) + 45(SWIR1) + 71(SWIR2)
MNDWI (Red)	(Red-MIR)/(Red+MIR)	GNDVI	(NIR-Green)/(NIR+Green)
MNDWI (Blue)	(Blue -MIR)/(Blue +MIR)	NDSI	(Green-SWIR)/(Green+SWIR)
NDVI	(NIR-Red)/(NIR+Red)	NDTI	(Red-Green)/(Red+Green)
NDMI	(NIR - SWIR)/(NIR + SWIR)	SWI	$1/\sqrt{(Blue - SWIR1)}$
NDPI	(SWIR - GREEN)/(SWIR + GREEN)		

B. Optimum Indices Selection

Feature selection methods have become an integral component of the machine learning processes for dealing with high-dimensional data. The selection of optimal features can be defined as the process of identifying relevant features that is able to obtain an optimal subset for well describing the problem and with minimal loss of efficiency [23]-[30]. Genetic algorithm uses Darwin's principles of natural selection to find the optimal way to predict or match the pattern. Genetic Algorithm is an evolutionary algorithm in which it is inspired by biology such as inheritance, mutation, natural selection and combination. Evolution starts from a completely random set of entities and is repeated in the next generations, and in each generation, the most appropriate solutions of the problem are selected.

The genetic algorithm was used in this research as the feature selection method in order to determine the optimal set of spectral indices for surface water bodies' detection and classification. As this algorithm works better for large set of features, in addition to the 19 spectral indices listed in Table 1, the spectral bands of Sentinel-2 satellite images were also used as features. Moreover, the support vector machine classifier was used as a cost function to evaluate the selected set of indices by the genetic algorithm for selecting the optimal set.

C. Evaluation by Decision Tree Classifier

The results of applying the SVM classifier on the optimal indices obtained from the genetic algorithm for detecting surface water bodies are compared with the results of applying the decision tree (DT) classifier for each individual spectral index in Table 1 in order to validate the optimization. The decision tree classifier places each pixel in a class by performing multi-stage classification using a series of binary decisions. The result of these successive decisions is a set of object classes. In this research, DT classifier has been used to classify pixels of each spectral index images into water or non-water land cover classes.

Results and Discussion

The pre-processing performed on satellite images in this research includes atmospheric corrections, normalization and cropping of images. QUick Atmospheric Correction (QUAC) algorithm is used for atmospheric correction of satellite images in ENVI software. This algorithm is suitable for correcting images with lack of the atmospheric and terrestrial samples. In this method, correction is done only by using several spectral bands and their wavelengths, and the image is corrected without providing additional information. Compared to other methods such as FLAASH, which are heavily affected by sensors' noise, the QUAC algorithm performs the correction process without affecting by the sensors' noise, and it also has very high speed.

The next pre-processing step is normalization in which, all data are placed between zero and one to have the same impact in performing measurements. Considering that the spectral indices are measured based on satellite images, the data should be normalized so that they are in the same range and very large or small values do not have a negative impact on the indices. After applying the pre-processing on the images those are taken from both study areas of Karkheh Dam and Shadegan wetland, all spectral indices are measured based on the formulas in Table 1. Then, the genetic algorithm is applied to determine the optimal set of these spectral indices. In order to improve the efficiency of the genetic algorithm, in addition to 19 indices in Table 1, 10 spectral bands of Sentinel-2 images were also added to the collection as individual spectral indices. These 10 spectral bands composed of three visible bands (B2, B3, B4), five NIR bands (B5, B6, B7, B8 and B8a) and two SWIR bands (B11 and B12).

The parameters setting of the Genetic algorithm are as follows: size of population=40, keeping rate=0.7, mutation probabilities=0.2 and maximum number of iterations=300. Table 2 shows the optimum set of spectral indices selected by the genetic algorithm based on SVM classification in each of the study areas of Karkheh Dam and Shadegan Wetland.

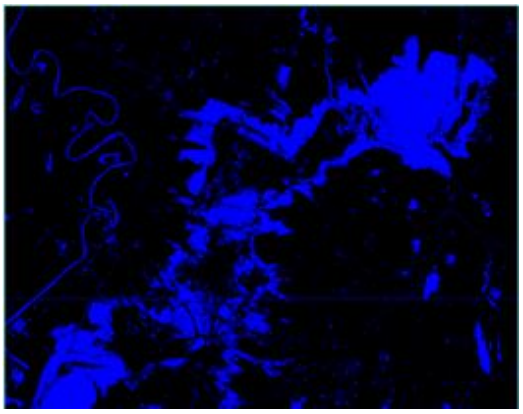
Table 2: Optimum spectral indices selected by genetic algorithm followed by SVM classifier

Study Area	Optimum Set of Indices			
Karkheh Dam	NDPI	MNDWI (Red)	AWEInsh	Band RedEdge3
Shadegan Wetland	NDPI	MNDWI	AWEIsh	SWI

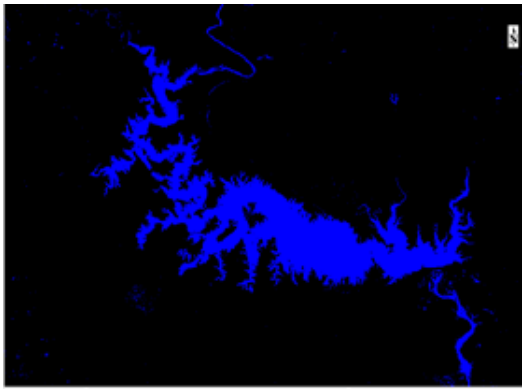
Fig. 3 shows the results of applying the SVM classifier using the optimum spectral indices selected by the genetic algorithm in both study areas.

In general, water absorbs the electromagnetic spectrum, and the reflection from the surface water bodies is insignificant compared to other land covers. The highest reflection from water bodies occurs in the blue band range, and the water's reflection reaches zero in the middle IR range of the electromagnetic spectrum. However, plants have the most absorption in the red band and the most reflection in the near infrared band.

According to the aforementioned interpretation of the spectral behavior of water and vegetation cover, the presence of Red, Red Edge, NIR and MIR bands in the formulas of the optimum indices selected by the genetic algorithm in the Karkheh dam area, which has relatively pure water, can be reasonable.



(a)



(b)

Fig. 3: SVM classification results using optimum indices in a) Shadegan wetland, b) Karkheh Dam.

These indices are able to detect and remove vegetation, soil and other land covers to identify pure surface water bodies.

But in the Shadegan wetland study area, surface water bodies are mixed with vegetation. Thus, the optimum selected indices by the genetic algorithm in this area have the blue band in their formulas, which has the highest reflection from the surface water.

As mentioned in the description of spectral indices, the AWEIsh index is suitable for use in areas with shadow effects. Considering that Shadegan wetland study area has water mixed with vegetation, the AWEIsh index has a better performance in this area due to its ability to detect water bodies from plants' shadows.

To evaluate the performance of the genetic algorithm in selecting the optimal set of spectral indices, the result of applying the SVM classifier on the optimal indices are compared with the results of applying the decision tree classifier on each individual index. Table 3 shows the overall accuracy and Kappa coefficient of the obtained results from applying decision tree classifier on each of the spectral indices in both study areas.

As it is depicted in Table 3, the highest accuracy of the decision tree classifier has been obtained in the Karkheh dam study area using the AWEInsh, NDWI and TCW2 spectral indices and in the Shadegan wetland study area using the AWEIsh, NDWI and GNDVI spectral indices.

The most obtained DT classification accuracy is 96.76% (by AWEInsh) in Karkheh dam and 96.48% (by AWEIsh) in Shadegan wetland, respectively. However, the SVM classifier based on the optimal set of selected indices (see Table 2) by the genetic algorithm has a higher overall accuracies and Kappa coefficients in both study areas.

For visual evaluation, the results of applying the DT classifier on each of the AWEInsh, NDWI and TCW2 spectral indices in the Karkheh Dam area (Fig. 4) and the AWEIsh, NDWI and GNDVI spectral indices in the Shadegan Wetland area (Fig. 5) have been compared with some other spectral indices.

In Fig. 6, the accuracy of the SVM classification results based on the optimal indices of the genetic algorithm is compared with the results of applying the decision tree classifier using AWEInsh in the Karkheh dam area and AWEIsh in the Shadegan wetland area.

Table 3: DT classification results of the individual spectral indices

Spectral Index	Study Areas			
	Karkheh Dam		Shadegan Wetland	
	Overall Accuracy (%)	Kappa	Overall Accuracy (%)	Kappa
WRI	90.51	0.87	93.60	0.87
WI2015	91.90	0.89	93.18	0.82
TCW2	**96.12**	**0.95**	84.26	0.56
TCW	91.24	0.88	91.47	0.72
SWI	94.27	0.89	95.77	0.91
NDWI	**96.72**	**0.95**	**96.26**	**0.91**
NDWI (RED)	89.04	0.76	93.61	0.79
NDWI (Blue)	92.66	0.88	93.66	0.85
NDVI	89.04	0.76	92.61	0.79
NDTI	92.60	0.89	87.68	0.95
NDSI	92.93	0.89	91.57	0.87
NDPI	95.70	0.92	95.57	0.85
NDMI	69.77	0.27	72.77	0.35
MNDWI	95.93	0.93	94.57	0.86
MNDWI (RED)	92.39	0.84	95.47	0.89
MNDWI (Blue)	93.33	0.90	95.14	0.89
GNDVI	90.88	0.82	**96.32**	**0.91**
AWEIsh	92.05	0.84	**96.48**	**0,90**
AWEInsh	**96.76**	**0.95**	82.31	0.52

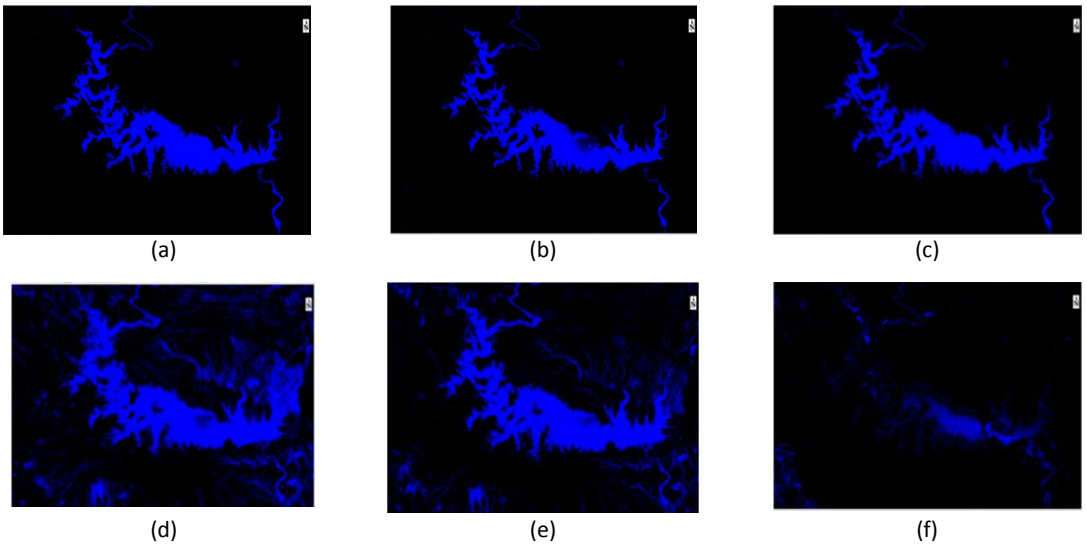


Fig. 4: Visual evaluation between the best decision tree classification results of Karkheh Dam by a) AWEInsh, b) NDWI and c) TCW2 indices compared with d) GNDVI, e) AWEIsh and f) NDMI.

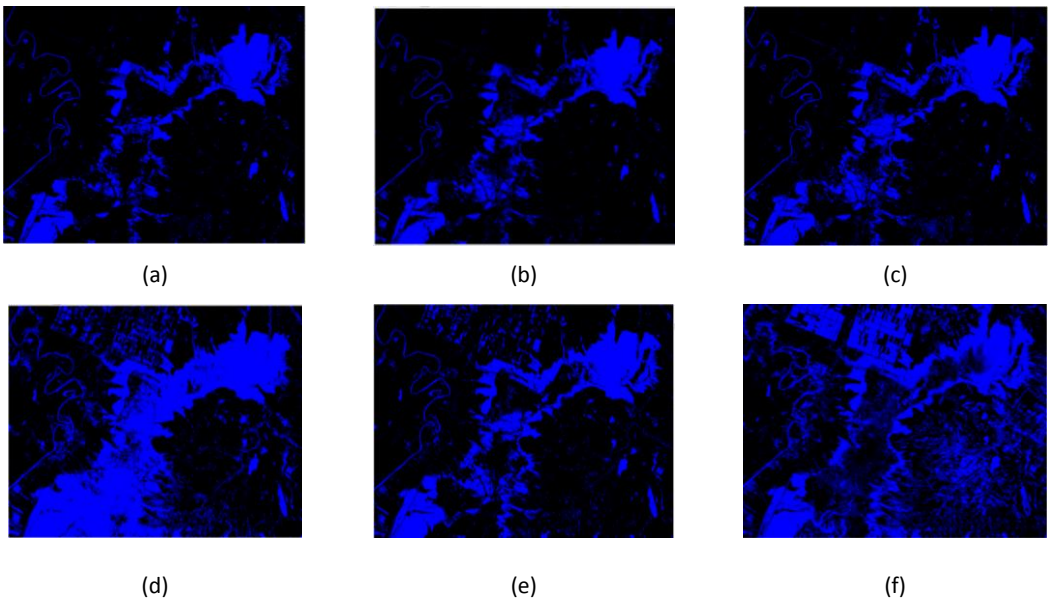


Fig. 5: Visual evaluation between the best decision tree classification results of Shadegan by a) AWEIsh, b) NDWI and c) GNDVI indices compared with d) AWEInsh, e) TCW and f) NDMI.

As it can be seen in Fig. 6, use of the optimal set of spectral indices selected by the genetic algorithm has been able to improve the classification accuracy of surface water bodies from other land covers.

Moreover, the obtained classification accuracies are compared with Random Forest classifier (Fig. 6). The comparison shows that random Forest classifier also has the overall accuracy 97.35 for Karkheh Dam and 97.14 for Shadegan wetland based on the optimum spectral indices by genetic algorithm in both study areas.

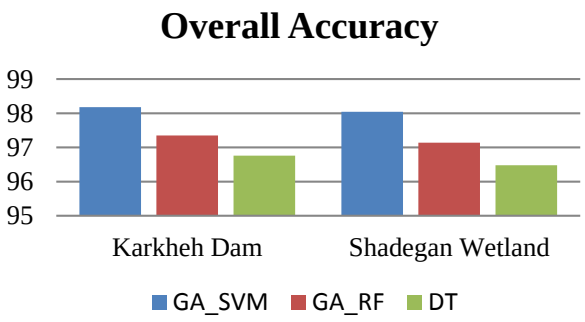


Fig. 6: Comparing the classification accuracies.

Conclusion

The main objective of this research is to evaluate the capabilities of the genetic algorithm to select an optimal set of spectral indices in order to classify surface water bodies from other land cover objects. For this purpose, after measuring the number of 19 conventional spectral indices in the detection of water bodies, each of these indices were classified into two classes of water and non-water objects using the decision tree classifier.

The evaluation of each of the spectral indices measured in both study areas of Karkheh Dam and Shadegan Wetland was carried out using quantitative decision tree classification criteria. Based on the differences in the nature of the two study areas (pure water of Karkheh Dam and water mixed with vegetation in Shadegan wetland), the number of three different indices in each area had the highest values of classification accuracy.

Applying the genetic algorithm on the spectral indices was able to obtain two optimal sets of effective indices that have the highest amount of accuracy in classifying water bodies from other land cover objects in each of the study areas. The evaluation of the obtained classification results of Sentinel-2 satellite images taken from the study areas showed that the use of optimum indices selected by the genetic algorithm could improve the overall classification accuracy by 1.42% in the Karkheh Dam area and 1.56% in the Shadegan Wetland area.

In addition to increasing the classification accuracy of the surface water bodies, considering the effects of the type of land cover objects on the efficiency of spectral indices, the use of genetic algorithm can significantly reduce the computational cost of comparing the results of a wide set of spectral indices. Moreover, considering the collective performance, genetic algorithm selects an optimal set of indices that can detect water bodies more accurately than any single index.

Funding

This work was supported by Shahid Rajaee Teacher Training University under grant number 4943.

Author Contributions

Hamzeh Karim Tabahfar has the following rolls in this manuscript: Data capturing, Formal analysis; Methodology implementation. Fatemeh Tabib Mahmoudi as the corresponding author has the following responsibilities: Supervision; Results' validation; Visualization; Writing - original draft; Writing - review & editing.

Conflict of Interest

The author declares that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

Acknowledgement

Authors acknowledge the dept. of Geomatics Engineering of Shahid Rajaee Teacher Training University for their supports.

Abbreviations

GA	Genetic Algorithm
DT	Decision Tree
NDWI	Normalized Difference Water Index
SVM	Support Vector Machine

References

[1] C. J. Vörösmarty, P. Green, J. Salisbury, R. B. Lammers, "Global water resources: Vulnerability from climate change and population growth," Science, 289(5477): 284-288, 2000.

[2] A. Karpatne, A. Khandelwal, X. Chen, V. Mithal, J. Faghmous, V. Kumar, "Global monitoring of inland water dynamics: State-of-the art, challenges, and opportunities," Studies in Computational Intelligence, 645: 121-147, 2016.

[3] L. Niu, H. Kaufmann, G. Xu, G. Zhang, C. Ji, Y. He, M. Sun, "Triangle Water Index (TWI): An advanced approach for more accurate detection and delineation of water surfaces in sentinel-2 data," Remote Sens., 14(21): 5289, 2022.

[4] L. Shen, C. Li, "Water body extraction from Landsat ETM+ imagery using adaboost algorithm," in Proc. 18th International Conference on Geoinformatics: 1-4, 2010.

[5] A. Saber, I. El-Sayed, M. Rabah, M. Selim, "Evaluating change detection techniques using remote sensing data: Case study New Administrative Capital Egypt," Egypt. J. Remote Sens. Space Sci., 24(2021): 635-648, 2021.

[6] J. Li, R. Ma, Z. Cao, K. Xue, J. Xiong, M. Hu, X. Feng, "Satellite detection of surface water extent: A review of methodology," Water, 14: 1148, 2022.

[7] J. Bhangale, S. More, T. Shaikh, S. Patil, N. More, "Analysis of surface water resources using sentinel-2 imagery," Procedia Comput. Sci. 171: 2645-2654, 2020.

[8] A. Ogilvie, G. Belaud, S. Massuel, M. Mulligan, P. Le Goulven, R. Calvez, "Surface water monitoring in small water bodies: potential and limits of multi-sensor Landsat time series," Hydrol. Earth Syst. Sci., 22: 4349-4380, 2018.

[9] S. K. McFeeters, "The use of normalized difference water index (NDWI) in the delineation of open water features," Int. J. Remote Sens., 17: 1425-1432, 1996.

[10] C. Kwang, E. M. Osei Jnr, A. S. Amoah, "Comparing of landsat 8 and sentinel 2a using water extraction indexes over volta river," J. Geogr. Geol., 10(1): 1-7, 2018.

[11] A. Fisher, N. Flood, T. Danaher, "Comparing Landsat water index methods for automated water classification in eastern Australia," Remote Sens. Environ. 175: 167-182, 2016.

[12] H. W. Khalid, R. M. Zahid Khalil, M. A. Qureshi, "Evaluating spectral indices for water bodies extraction in western Tibetan Plateau," Egypt. J. Remote Sens. Space Sci. 24(2021): 619-634, 2021.

[13] D. Montero, C. Aybar, M. D. Mahecha, F. Martinuzzi, M. Söchting, S. Wieneke, "A standardized catalogue of spectral indices to advance the use of remote sensing in Earth system research," Sci. Data, 10(197), 2023.

[14] T. D. Acharya, A. Subedi, I. T. Yang, D. H. Lee, "Combining water indices for water and background threshold in landsat image," In Multidisciplinary Digital Publishing Institute Proceedings, 2(3): 143, 2017.

[15] G. Feyisa, H. Meilby, R. Fensholt, S. Proud, "Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery," Remote Sens. Environ., 140: 23-35, 2014.

[16] H. Emami, A. Zarei, "Modeling lake water's surface changes using environmental and remote sensing data: A case study of lake Uremia," Remote Sens. Appl. Soc. Environ., 23: 100594, 2021.

[17] L. Ji, Zhang, B. Wylie, "Analysis of dynamic thresholds for the normalized difference water index," Photogram. Eng. Remote Sens., 75: 1307-1317, 2009.

[18] B. P. Salmon, W. Kleynhans, F. Van Den Bergh, J. Olivier, T. L. Grobler, K. J. Wessels, "Land cover change detection using the internal covariance matrix of the extended Kalman filter over multiple spectral bands," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 6(3): 1079-1085, 2013.

[19] M. Volpi, G. P. Petropoulos, M. Kanevski, "Flooding extent cartography with Landsat TM imagery and regularized kernel Fisher's discriminant analysis," Comput. Geosci., 57 : 24-31, 2013.

[20] C. Huang, Y. Chen, S. Zhang, J. Wu, "Detecting, extracting, and monitoring surface water from space using optical sensors: A review," Rev. Geophys., 56: 333-360, 2018.

[21] H. Q. Xu, "Modification of normalized difference water index (NDWI) to enhance open water features in remotely sensed imagery," Int. J. Remote Sens., 27(14): 3025-3033, 2006.

[22] A. Fisher, T. A. Danaher, "Water index for spot5 hrg satellite imagery, New South Wales, Australia, Determined By linear discriminant analysis," Remote Sens., 5(11): 5907-5925, 2013.

[23] M. G. Altarabichi, S. Nowaczyk, S. Pashami, P. Sheikholharam Mashhadi, "Fast genetic algorithm for feature selection-A qualitative approximation approach," Expert Syst. Appl., 211: 118528, 2023.

[24] B. Olueye, A. Leisa, J. Leng, D. Dean, "A genetic algorithm-based feature selection," Int. J. Electron. Commun. Comput. Eng., 5(4): 899-905, 2014.

[25] F. Samadzadegan, F. Tabib Mahmoudi, "Optimum band selection in hyperspectral imagery using swarm intelligence optimization algorithms," in Proc. International Conference on Image Information Processing (ICIIP), 2011.

[26] F. Tabib Mahmoudi, "Investigation of water stress status of plants in north of Iran under under the influence of quarantine quarantine application in Covid-19 virus pandemic," J. Water Soil Conserv., 27(6), 2021.

[27] F. Tabib Mahmoudi, "Semantic object-based urban scene analysis for feature fusion of VHR imagery and Lidar DSM" Signal, Image Video Process., 2022.

[28] S. Khoramak, F. Tabib Mahmoudi, "Multi-agent hyperspectral and lidar features fusion for urban vegetation mapping," Earth Sci. Inf., 2023.

[29] O. Kavats, D. Khramov, K. Sergieieva, "Surface water mapping from SAR images using optimal threshold selection method and reference water mask," Water, 14: 4030, 2022.

[30] N. Nasir, A. Kansal, O. Alshaltone, F. Barneih, A. Shanableh, M. Al-Shabi, A. Al Shammaa, "Deep learning detection of types of water-bodies using optical variables and ensembling," Intell. Syst. Appl., 18: 200222, 2023.

Biographies

Hamzeh Karim Tabahfar received his B.Sc. degree in Civil Engineering the branch of Geomatics (Surveying and mapping) from Khuzestan University, Khuzestan, Iran, in 2019. Since 2020, he is the M.Sc. student of Remote Sensing in Geomatics department of the faculty of Civil Engineering, Shahid Rajaee Teacher Training University.

- Email: hamzehtabbahfar@mail.ir
- ORCID: 0009-0009-3906-7409
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Fatemeh Tabib Mahmoudi received her B.Sc. degree in Civil Engineering the branch of Geomatics (Surveying and mapping) from Khajeh Nasiredin Toosi University, Tehran, Iran, in 2004. She received his M.Sc. and PhD degrees in Photogrammetry from Tehran University, Tehran, Iran, in 2009 and 2014, respectively. Since 2016, she has been working as an assistant professor in the Geomatics department of the faculty of Civil Engineering, Shahid Rajaee Teacher Training University. She has some publications in the field of remote sensing data analysis, pattern recognition and data fusion.

- Email: fmahmoudi@sru.ac.ir
- ORCID: 0000-0002-8414-8189
- Web of Science Researcher ID: NA
- Scopus Author ID: 36669646900
- Homepage: NA

How to cite this paper:
H. Karim Tabahfar, F. Tabib Mahmoudi, "Optimum spectral indices for water bodies recognition based on genetic algorithm and sentinel-2 satellite images," J. Electr. Comput. Eng. Innovations, 12(1): 217-226, 2024.

DOI: 10.22061/jecei.2023.10118.678

URL: https://jecei.sru.ac.ir/article_2001.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A New Approach to Synthesis of a Quasi Non-Uniform Leaky Wave Antenna

A. Kiani, F. Geran *, S. M. Hashemi

Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:
Received
Reviewed
Revised
Accepted

Keywords:
leaky wave antenna (LWAs)
Quasi Uniform LWA
Quasi Non ULWA
Waveguide antenna

*Corresponding Author's Email Address: *F.geran@sru.ac.ir*

Abstract

Background and Objectives: In this paper, a closed-form mathematical formula has been presented using of the proposed periodic structure E-field distribution, that helps designers to calculate the width of the slots in Quasi Non-Uniform Leaky Wave Antenna (QNULWA).

Method: This method is based on two steps. In the first step, some important parameters for the proposed antenna design will be extracted using simulation. In the second step, by solving a discrete differential equation, a general relation will be obtained for these types of antennas. This method has been investigated in the case of slot LWA families.

Results: A Leaky wave antenna has been synthesized in the 15.5- 18 GHz frequency range for Gaussian radiation pattern. The results of simulation and antenna design will be very close to each other in 2.5 GHz Bandwidth (15.5 - 18 GHz), which shows the accuracy of this formula. Also, By changing the frequency range *2.5* GHz, the main lobe direction of the antenna will scan the space approximately *10* degrees (from 63 to 73 degree). The antenna has an SLL value of about *-25* dB and *13* dB Gain at whole band 15.5-18 GHz.

Conclusion: The obtained formula helps the antenna designers to calculate the dimensions this type of antenna for any pattern distribution.

Introduction

A leaky wave antenna is a type of antenna that radiates electromagnetic waves along its length rather than at its ends. It is called "leaky" because it allows some energy to leak out along the length of the antenna, resulting in radiation. Really, these antennas are designed to have a controlled radiation pattern along their length. This means that the direction and shape of the radiated beam can be adjusted by changing the properties of the antenna. By controlling the leakage of energy along the antenna, it is possible to steer the beam in a desired direction [1]-[5]. There are several types of leaky wave antennas, including:

1. Slotted Waveguide Antennas [6]-[8]: These antennas consist of a waveguide with slots cut into its walls. The slots allow energy to leak out and form a radiating wave along the length of the waveguide.

2. Dielectric Slab Leaky Wave Antennas [9]-[12]: These antennas use a dielectric slab placed on top of a ground plane. By varying the thickness and permittivity of the dielectric slab, the radiation pattern can be controlled.

3. Printed Leaky Wave Antennas: These antennas are made using printed circuit board technology. They consist of a microstrip transmission line with slots or periodic structures that allow energy to leak out and form a radiating wave.

4. Metamaterial Leaky Wave Antennas [13]-[17]: These antennas use metamaterials, which are artificially engineered materials with unique electromagnetic properties. By designing the metamaterial structure, it is possible to control the radiation pattern of the antenna.

Overall, leaky wave antennas offer flexibility in controlling the direction and shape of the radiated beam, making them suitable for various applications such as radar systems, wireless communication, and satellite communication.

No matter which category the leaky wave antenna is in, another category can be defined for it, which will be explained in more detail below.

A) A uniform leaky wave antenna is designed to have a consistent and constant radiation pattern along its length. It consists of a waveguide or transmission line with regularly spaced slots or apertures. The electromagnetic waves leak out gradually through these slots or apertures, resulting in a uniform radiation pattern without any significant variation in the radiation intensity along the antenna's length.

B) A non-uniform leaky wave antenna, on the other hand, is designed to have a varying radiation pattern along its length. It consists of a waveguide or transmission line with non-uniformly spaced slots or apertures. The non-uniform spacing of these slots or apertures causes the electromagnetic waves to leak out at different rates along the length of the antenna. As a result, the radiation pattern of the antenna varies along its length, allowing for controlled variation in the direction and intensity of the radiated waves.

C) A quasi-uniform leaky wave antenna is a combination of both uniform and non-uniform leaky wave antennas. It consists of a waveguide or transmission line with a combination of regularly spaced and non-uniformly spaced slots or apertures. This design allows for a partially uniform and partially non-uniform radiation pattern along the length of the antenna. The quasi-uniform leaky wave antenna provides some flexibility in directing and shaping the radiated waves while still maintaining a relatively consistent radiation pattern.

D) Lastly, a quasi-non-uniform leaky wave antenna is also a combination of uniform and non-uniform leaky wave antennas. It consists of a waveguide or transmission line with a combination of regularly spaced and non-uniformly spaced slots or apertures. However, in this design, the non-uniform spacing is more dominant, resulting in a radiation pattern that is mostly non-uniform with some elements of uniformity. The quasi-non-uniform leaky wave antenna offers a compromise between the constant radiation pattern of a uniform antenna and the variable radiation pattern of a non-uniform antenna, providing flexibility in directing and shaping the radiated waves.

Overall, uniform/ non-uniform/ quasi-uniform/ quasi-non-uniform leaky wave antennas offer unique advantages depending on the specific application requirements. The choice between them depends on factors such as desired radiation pattern, beam-width, and scanning capabilities.

By examining the research done in this field, we can find valuable papers, some of which deal with the synthesis and others with the design of this type of antenna. Among others, we can mention the relationship presented by Oliner [18], which is of particular importance in the synthesis of the shape of the tapering slot in the leaky wave antenna [18]-[20]. The disadvantages of this design include the ability to scan the space in only the forward quadrant space of the antenna and the difficult implementation to create the exact tapering slot. To overcome these problems, the periodic LWAs were introduced by researchers. Indeed, periodic discrete slots are used instead of continuous slot. These antennas can scan the backward and forward quadrant space of the antenna [21]-[24].

Of course, in addition to the ability to scan one or two quarters of the sphere by the leaky wave antenna, increasing the scanning angle with low frequency changes, the stability of the radiation pattern with changing the scanning angle are also the challenges of this type of antenna design, which different designs have tried to overcome this challenge [25]-[27].

The following paper is a continuation and completion of a paper [28] in which the structure of the slots and design has changed. These changes lead to the creation of a new structure called QNU-LWA, where the width of the slots have been modified along antenna and this causes the distance between the slots to change and leads to QNU-LWA. Fig. 1 shows the difference between these two structures. In the structure of QU-LWA, the phase of the E-field distribution in each slot is fixed and their amplitude is changing, but in the structure of QNU-LWA in addition to the amplitude, the phase of the E-field distribution is changing in each slot.

In paper [28], considering that the distance between the slots were equal and only the length of the slots was different, a mathematical formula was presented to calculate the length of each slot according to the desired radiation pattern. While in this paper, the length of the slots is fixed and the width of the slots are different, which is the difference between these two structures. As a result, the main innovation of this paper is to present a mathematical formula for the distribution of the desired radiation pattern to calculate the slots width of the proposed structure. Finally, to confirm the obtained mathematical formula, an antenna sample with Gaussian radiation pattern was designed and constructed in the frequency range of 15.5 to 18 GHz.

This paper consists of the following sections: In Section II, the method of obtaining the important information required for the synthesis of these antennas will be mentioned. In Section III, the mathematical design

method of the antenna will be explained. In Section IV, the simulation and construction results of this antenna will be compared. Finally, Section V concludes the paper.

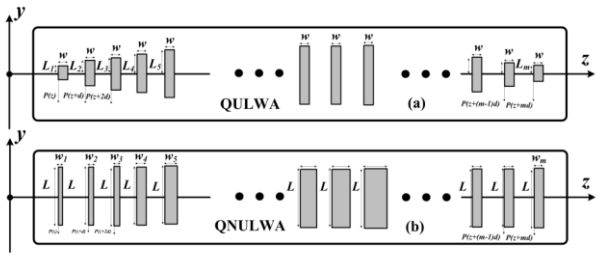


Fig. 1: (a) QU-LWA and (b) QNU-LWA model.

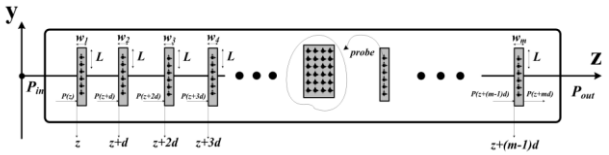


Fig. 2: Simulated antenna model to obtain the E-field distribution inside the slots.

Initial Step for Synthesizing QNU-LWA

To design this antenna, in the first step drew a QU-LWA structure in the CST software based on Fig. 2. The mentioned method in this section can be generalized to other types of LWA [28]. This structure is just an example and draws only to get the mathematical relations for the electric field inside the slots of this antenna. After the mathematical relation for antenna design is obtained in the next section, there is no need to do the first step in antenna design and can be ignored. A series of probes to measure the E-field inside these slots will then be placed on each slot. In this structure, w_m and L are the width and length of the slots and are considered constant. m is the number of antenna slots and is selected based on the $S_{21} = -10\ dB$ dispersion parameter. There are two ways to achieve $S_{21} = -10\ dB$. The first method is to increase the number of antenna slots. In this case, as shown as an example in Fig. 3, S_{21}will decrease by increasing the value of the parameter m. One of the problems with this method is the increased antenna length and the process of manufacturing and machining the antenna will be difficult. On the other hand, by increasing the length of the antenna, it will be easier to achieve a narrower 3 dB pattern. Fig. 3 shows the increase in the number of slots and the effects of this increase on the S_{21} parameter. As expected, by increases the number of slots, the value of S_{21} will decrease.

In the second method, the width of each slot w increases. As shown in Fig. 4, by increasing the w, the leakage field of each slot to the outside will increase and as a result, the parameter S_{21} will decrease.

One of the problems of this method is increasing the 3 dB beamwidth of the antenna. In antenna design, based on the need for gain and 3 dB beamwidth and antenna length limit, we select this parameter to access the appropriate S_21 value. By selecting the appropriate value for m, the distribution of the E-field inside the slot is drawn in Fig. 5. This figure shows an exponential function whose amplitude decreases along whit the antenna. This figure will change very little if the width of the slots or the number of slots changes. This figure will be the basis of antenna design in the next step.

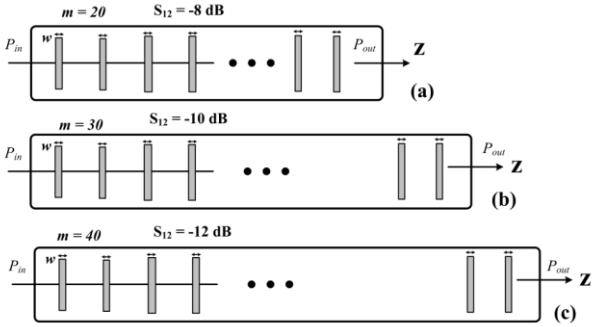


Fig. 3: Increase the number of slots and the effects of this increase on the S_{12} value.

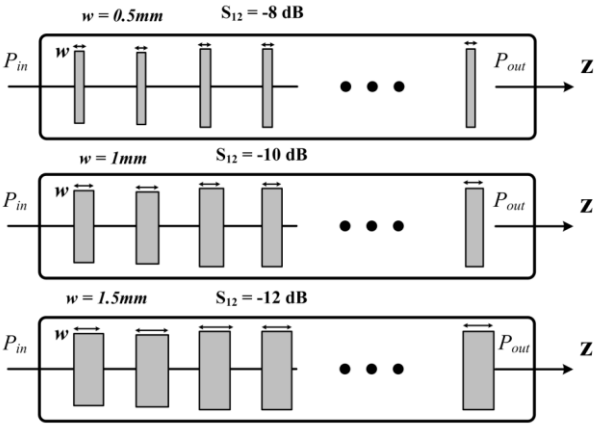


Fig. 4: Increase the width of each slots and the effects of this increase on the S_{12} value.

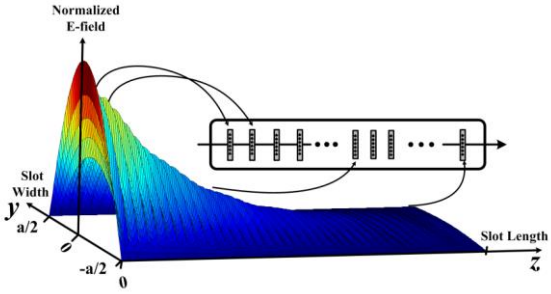


Fig. 5: The E-field distribution inside each slot.

Method of Extraction of the Mathematical Formula

According to distribution of the E-field drawn in Fig. 5, it can be written for the E-field inside the antenna slot as follows [29]-[30]:

$$E_y = \sum_{n=-\infty}^{n=\infty} E_0 cos\left(\frac{\pi}{a}y\right) exp(-\alpha z) exp\left(-j\left(\beta + \frac{2n\pi}{d}\right)z\right) \quad (1)$$

where n th Floquet mode or also referred to as space harmonic. d indicates the fixed period of each slot and a is waveguide width. α and β are leakage and phase constant of the waveguide respectively. The value of E_x and E_z is very small compared to E_y component and can be neglected. Based on the Floquet's theorem and the 3D distribution of the E-field in Fig. 5, it is expected that the E-field and the radiation power are periodic. In Fig. 5, function $exp\left(-j\left(\beta + \frac{2n\pi}{d}\right)z\right)$ decreases with the exponential function $exp(-\alpha z)$ during the antenna length. First, it is supposed that P_{in} is the input power applied to the antenna, some input power is radiated by each slot and the remained power at the antenna end is absorbed by the match load. w_1, w_2, ..., w_i and L are the width and length of the slots of each segment. In the QU-LWA antenna design, the amount of w is constant and L is changed in each slot, while in QNU-LWA, L is constant and w changes in each slot. Based on the periodic structure in Fig. 5 and Floquet's theorem, geometric periodicity forces the field to be periodic. So, it can be defined as a periodic function for the radiation power of each segment $P(z)$ as [13]:

$$\begin{aligned} &P(z+d) = CP(z) \\ &P(z+2d) = CP(z+d) \\ &P(z+3d) = CP(z+2d) \\ &\vdots \\ &P(z+id) = CP(z+(i-1)d), i = 1,2,\dots,m \end{aligned} \quad (2)$$

where C^i is the form of:

$$C^i = \sum_{n=-\infty}^{n=\infty} \exp(-2\alpha i d) \exp\left(-2j\left(\beta + \frac{2n\pi}{d}\right) i d\right) \quad i = 1,2,\dots,m \quad (3)$$

where m is the number of slots in the antenna. In (3), $P(z)$ is the applied power to the antenna. Part of the power will be propagated to the outside of the antenna and $P(z+id)$ is applied power to the i th slot, where $i = 1,2,\dots,m$. The radiation power in each slot is proportional to the input power, size of the slot, and $|E_y|^2$ coefficient. So, for total radiation in each slot, it can be written based on [29] as follows:

$$P_{rad} = \frac{P_{in} - P_{in}}{P_{out}} = cos^2(\pi/a\, L_i) A \times L \sum_{i=1}^{m} w_i C^i = \sum_{i=1}^{m} P_{r,i} \quad (4)$$

A is a coefficient of $1/meter^2$ that will be calculated in the future. $P_{in} = P(z)$ is input power, $P_{out} = P(z+id)$ is output power and $y = L$. While P_{rad} is the total radiation power of the antenna and $P_{r,i}$ is the radiation power in each slot. In order to implement a mathematical distribution function, the radiation power must follow the amplitude distribution such as Gaussian. Suppose that the desired aperture distribution or slot radiation is T_i, $i = 1,2,\dots,m$, so:

$$cos^2(\pi/a\, L) A \times L \times w_i C^{i-1} = BT_i^2 \quad (5)$$

where B is a constant coefficient. Finally, gives w_i the form of (6):

$$w_i = \frac{BT_i^2}{A \times L \times cos^2(\pi/a\, L) C^{i-1}} \quad (6)$$

One of the unknown parameter of (6) is A. To calculate A according to $i = 1$ in the (5), $BT_1^2 = A \times L \times cos^2(\pi/a\, L) w_1$, it can be written as:

$$A = \frac{BT_1^2}{w_1 \times L \times cos^2(\pi/a L)} \quad (7)$$

Since, in (6), the amplitude distribution is calculated for a slot on the antenna, the value of C^i should be an absolute value and therefore used $|C^i|$ instead of C^i. Combining (6) and (7), it can be concluded that w_i as:

$$w_i = \frac{w_1}{|C^{i-1}|}\left(\frac{T_i}{T_1}\right)^2, w_i < d, \quad i = 1,2,\dots,m \quad (8)$$

Equation (8) is a very simple relation that shows, to calculate the width of the slots, only needed the width of the initial slot w_1, C^i and distribution coefficients T^i. To get the initial conditions w_1 consideration the $T_m^2 = T_1^2$ for distribution coefficients, so $w_1 < |C^{m-1}|d$ and as a result, the allowable range for w_1 is shown with dashed line in Fig. 6.

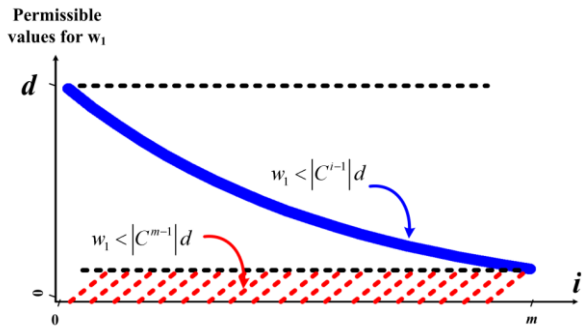


Fig. 6: Allowable range for selection w_1 parameter.

By selecting a suitable interval for w_1 and using (8), the w_i coefficients can be obtained as Table 1. In this case, Gaussian distribution coefficients are used. Other distribution coefficients such as ChebyChef, Taylor, and ... can also be used and the different properties of these distributions can be implemented on the antenna. In this structure, the distance between the two slots d is fixed, and the width of the slot w_1 is changing based on Table 1. By selecting the Gaussian distribution and using (8), the

slot shape on the antenna wall can be designed. Fig. 7 shows how the use of a Gaussian distribution leads to the creation of a series of slot structures on the antenna wall.

Table 1: The value of calculated w_i

i	w_i	i	w_i	i	w_i	i	w_i
1	0.5745	11	1.8366	21	3.4211	31	3.8431
2	0.6649	12	1.9960	22	3.5539	32	3.7151
3	0.7640	13	2.1583	23	3.6750	33	3.5494
4	0.8719	14	2.3223	24	3.7817	34	3.3493
5	0.9882	15	2.4870	25	3.8711	35	3.1203
6	1.1128	16	2.6512	26	3.9398	36	2.8696
7	1.2450	17	2.8138	27	3.9841	37	2.6056
8	1.3842	18	2.9735	28	4.0000	38	2.3367
9	1.5298	19	3.1290	29	3.9839	39	2.0707
10	1.6809	20	3.2788	30	3.9324	40	1.8145

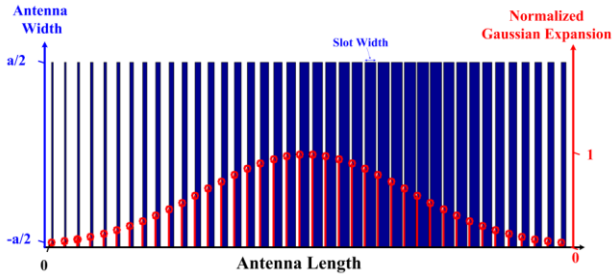


Fig. 7: The Width of the slots designed based on (8) and the Gaussian coefficients to be created on the antenna wall.

Simulations and Results

For a detailed study of the mathematical relationship expressed in (8), compared in this section simulations results with practical measurement results. The photograph of the construction antenna is shown in Fig. 8. The antenna consists of two parts that are connected with screws. The lower part is made of aluminum and the upper part is made of stainless steel to be able to withstand machining stresses. Due to the need for high accuracy in the manufacturing process, the CNC machine with micron precision has been used. Making the first slots $(i = 1 - 2, w_1 = 0.5645,\ w_2 = 0.6649)$ were complicated because they were so thin and it is with some errors.

The manufacturing antenna parameter is shown in Table 2. The simulation and measure test results pattern of the antenna are shown in Fig. 9. The antenna at 2.5 GHz bandwidth shows very acceptable results compares to the simulation. By changing the frequency range 2.5 GHz, the main lobe direction of the antenna will scan the space approximately 10 degrees (from 63 to 73 degree). The antenna has an SLL value of about -25 dB and 13 dB Gain. By the theory of Gaussian expansion coefficients with coefficient number 40 and $\sigma = 1.7$, we should reach SLL about -27 dB. Therefore, the results of simulation and antenna design are very close to the theoretical results. Fig. 10 shows S_{11} and S_{21} in both simulation and measure test. The results of the simulation and measure tests are very similar. An important point is that initially the antenna was designed for $S_{21} = -10$ dB, but in Fig. 10 it has reached about $S_{21} = -40$ dB. The reason for this difference is that we initially chose the width of the slots as fixed, but changed the width of the slots using (8), These changes have improved the antenna radiation to the outside space and the value of S_{21}has been somewhat improved.

Table 2: Antenna physical parameter

a	b	L_{slot}	L_{total}	t	d	m	f_0	BW
10.1 mm	5 mm	175 mm	235 mm	1 mm	3.8 mm	40	16.75 GHz	2.5 GHz

Fig. 8: The photograph of the fabricated antenna.

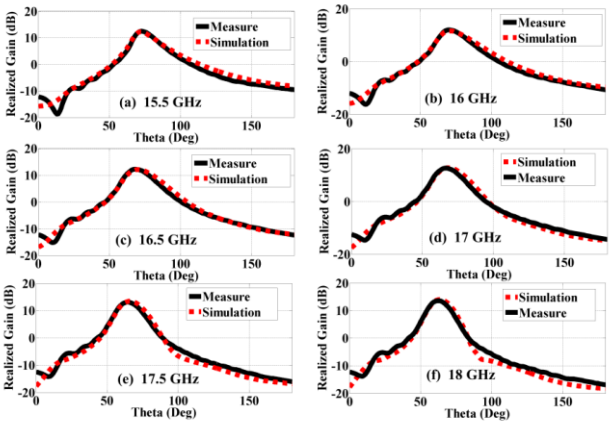


Fig. 9: Theoretical and measured patterns for the QNULWAs. (a) 15.5 GHz. (b) 16 GHz. (c) 16.5 GHz. (d) 17 GHz. (e) 17.5 GHz. (f) 18 GHz. based on (8) and The value of calculated w_i (Table 1).

Variations of the beam angle and α/k_0 based on the variation of frequency are shown in Fig. 11. The antenna main beam scans with frequency changes; in fact, it moves from 64 to 72 degree as the frequency decrease from 18 to 15.5 GHz. Through analyzing the changes in α/k_0 and β/k_0, caused by frequency variations, the possible radiation frequency range can be found. In other words, the dimensions for the structure, which make the desired radiation pattern possible within the operating frequency range, are selected.

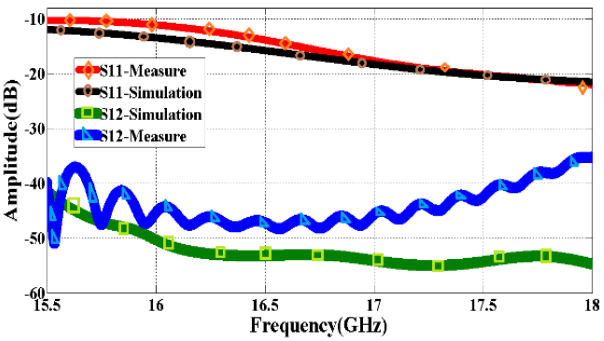


Fig. 10: Theoretical and measurement S_{11} and S_{12}.

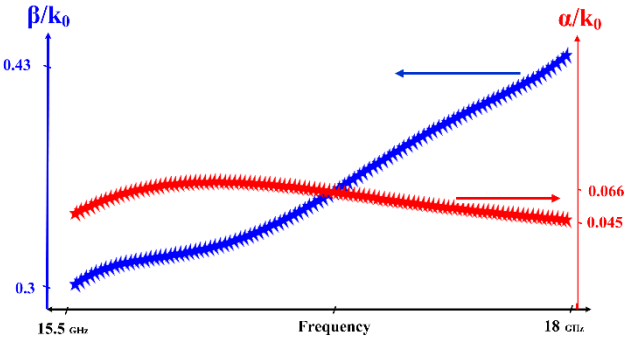


Fig. 11: Leaky-mode propagation constant $\boldsymbol{\alpha/k_0}$ and $\boldsymbol{\beta/k_0}$.

As mentioned in the introduction, the main goal of this paper is to derive the mathematical formula for QNU-LWA antenna synthesis. Then, an antenna sample was designed to confirm the formula. Finally, Table 3 compares the design antenna specifications with some other references reported in recent years.

Table 3: Compare of the design antenna specifications with some other references

Reference no.	Frequency range (GHz)	Scanning range (Degree)	Gain (dB)	Side-Lob level (dB)	Antenna Type
[22]	11.7 to 19.6	-61 to +34	14.1	-12	A symmetrical SIW
[23]	9.5-13.7	+5 to +81	15.7	-15	SIW with Dumbbell-shaped slots
[24]	9.3 to 9.93	-65 to +65	NA	NA	Composite right/left-handed
[25]	6 to 16	-68 to +23	16.86	NA	Hole array spoof surface plasmon polaritons
[26]	9.2 to 10.8	58 to 62	14.5	-20	Quasi-uniform leaky-wave antenna
This work	15.5 to 18	63 to 73	13	-25	Quasi-Non-Uniform leaky-wave antenna

Results and Discussion

In this paper, a mathematical formula to synthesis the QNU-LWA was presented. Using a method based on simulation and obtaining the distribution E-field inside the slots and solving a differential equation, it was possible to calculate the width of slots in the QNU-LWA.

To verify the obtained formula, an antenna with Gaussian distribution for the radiation pattern was synthesized and constructed. The simulation results confirm the correctness of the formula. Also, the simulation and construction results of this antenna are exactly the same.

Author Contributions

A. Kiani and F. Geran conceived the idea, analyzed the theoretical feasibility, and wrote the manuscript. S.M. Hashemi carried out the full-wave simulations and performed the measurements.

Acknowledgment

The authors would like to thank the anonymous reviewers and the editors of JECEI for their valuable comments and suggestions for improving quality of the paper.

Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

QU-LWA	Quasi Uniform Leaky Wave Antenna
QNU-LWA	Quasi non-uniform leaky wave antenna

References

[1] V. Arya, T. Garg, "Leaky wave antenna: A historical development," Microwave Rev., 27(1): 3-16, 2021.

[2] D. Jackson, P. Baccarelli, P. Burghignoli, W. Fuscaldo, A. Galli, G. Lovat, "A history of leaky waves and leaky-wave antennas," in Proc. IEEE 2019 URSI International Symposium on Electromagnetic Theory (EMTS): 1-4, 2019.

[3] D. R. Jackson, C. Caloz, T. Itoh, "Leaky-wave antennas," Proc. IEEE, 100(7): 2194-2206, 2012.

[4] M. K. Mohsen et al., "The fundamental of leaky wave antenna," J. Telecommun., Electr. Comput. Eng. (JTEC), 10(1): 119-127, 2018.

[5] F. Monticone, A. Alu, "Leaky-wave theory, techniques, and applications: From microwaves to visible frequencies," Proc. IEEE, 103(5): 793-821, 2015.

[6] F. L. Whetten, C. A. Balanis, "Meandering long slot leaky-wave waveguide-antennas," IEEE Trans. Antennas Propag., 39(11): 1553-1560, 1991.

[7] A. R. Mallahzadeh, M. H. Amini, "Design of a leaky-wave long slot antenna using ridge waveguide," IET Microwaves Antennas Propag., 8(10): 714-718, 2014.

[8] S. K. Lin, Y. C. Lin, "A compact outer-fed leaky-wave antenna using exponentially tapered slots for broadside circularly polarized radiation," IEEE Trans. Antennas Propag., 60(6): 2654-2661, 2012.

[9] T. Teshirogi, Y. Kawahara, A. Yamamoto, Y. Sekine, N. baba, M. Kobayashi, "High-efficiency, dielectric slab leaky-wave antennas," IEICE Trans. Commun., 84(9): 2387-2394, 2001.

[10] S. Mohanty, B. Mohapatra, "Leaky wave-guide based dielectric resonator antenna for millimeter-wave applications," Trans. Electr. Electron. Mater., 22(3): 310-316, 2021.

[11] S. N. Tcvetkova, E. Martini, S. A. Tretyakov, S. Maci, "Perfect conversion of a TM surface wave into a TM leaky wave by an isotropic periodic metasurface printed on a grounded dielectric slab," IEEE Trans. Antennas Propag., 68(8): 6145-6153, 2020.

[12] D. Jackson, A. Oliner, A. Ip, "Leaky-wave propagation and radiation for a narrow-beam multiple-layer dielectric structure," IEEE Trans. Antennas Propag., 41(3): 344-348, 1993.

[13] C. Caloz, T. Itoh, A. Rennings, "CRLH metamaterial leaky-wave and resonant antennas," IEEE Antennas Propag. Mag., 50(5): 25-39, 2008.

[14] S. Lim, C. Caloz, T. Itoh, "Metamaterial-based electronically controlled transmission-line structure as a novel leaky-wave antenna with tunable radiation angle and beamwidth," IEEE Trans. Microwave Theory Tech., 52(12): 2678-2690, 2004.

[15] K. M. Kossifos, M. A. Antoniades, "A NRI-TL metamaterial leaky-wave antenna radiating at broadside with zero beam-squinting," IEEE Antennas Wireless Propag. Lett., 17(12): 2223-2227, 2018.

[16] S. Lim, C. Caloz, T. Itoh, "Metamaterial-based electronically controlled transmission-line structure as a novel leaky-wave antenna with tunable radiation angle and beamwidth," IEEE Trans. Microwave Theory Tech., 52(12): 2678-2690, 2004.

[17] M. Alibakhshikenari et al., "Beam-scanning leaky-wave antenna based on CRLH-metamaterial for millimetre-wave applications," IET Microwaves Antennas Propag., 13(8): 1129-1133, 2019.

[18] T. Tamir, A. A. Oliner, "Guided complex waves. Part 1: Fields at an interface," in Proc. the Institution of Electrical Engineers., 110(2): 310-324, 1963.

[19] T. Tamir, A. A. Oliner, "Guided complex waves. Part 2: Fields at an interface," in Proc. The Institution of Electrical Engineers., 110(2): 325-334, 1963.

[20] S. Mohammad-Ali-Nezhad, A. Mallahzadeh, "Periodic ridged leakywave antenna design based on siw technology," IEEE Antennas Wireless Propag. Lett., 14: 354-357, 2014.

[21] D. K. Karmokar, K. P. Esselle, "Periodic U-Slot-loaded dual-band half-width microstrip leaky-wave antennas for forward and backward beam scanning," IEEE Trans. Antennas Propag., 63: 5372 - 5381, 2015.

[22] P. Sohrabi, P. Rezaei, S. Kiani, M. Fakhr, " A symmetrical SIW-based leaky-wave antenna with continuous beam scanning from backward-to-forward through broadside," Wireless Networks, 27(8): 5417–5424, 2021.

[23] A. Abolfathi, P. Rezaei, M. Sharifi, "Compact bilayer substrate integrated waveguide leaky wave antenna with dumbbell-shaped slot based on the TE 20 mode," Int. J. RF Microwave Comput. Aided Eng., 29(8): e21791, 2019.

[24] W. Yang, Z. Peng, J. Ren, J. Gao, G. Zhai, "Composite right/left-handed leaky-wave antenna with high scanning rate," IEEE Antennas Wireless Propag. Lett., 21(12): 2522-2526, 2022.

[25] J. Wang, K. Xu, X. Kong, R. Xu, L. Zhao, "Wide-Angle beam scanning leaky-wave antenna array based on hole array SSPPs," IEEE Antennas Wireless Propag. Lett., 22(7): 1731-1735, 2023.

[26] N. Javanbakht, M. S. Majedi, A. R. Attari, "Thinned array inspired quasi-uniform leaky-wave antenna with low side-lobe level," IEEE Antennas Wireless Propag. Lett., 16: 2992-2995, 2017.

[27] A. Kiani, F. Geran, S. M. Hashemi, K. Forooraghi, "A presentation of a mathematical formula to design of a quasi-uniform leaky-wave antenna with ultra-low side-Lobe level," IEEE Antennas Wireless Propag. Lett., 18: 901-905, 2019.

[28] A. Kiani, F. Geran, S. M. Hashemi, K. Forooraghi, "Mathematical analysis of a modified closed-form formula for design a uniform leaky wave antenna with ultra-low SLL," Sci. Rep., 9: 9372, 2019.

[29] C. A. Balanis, Modern antenna handbook, John Wiley and Sons, 2008, ch. 7.

[30] A. A. Oliner, D. R. Jackson, Leaky-wave antennas, in Antenna Engineering Handbook., 4th ed., J. L. Volakis, Ed. New York, NY, USA: McGraw-Hill, 2007, ch. 11.

Biographies

Abdoreza Kiani was born in Khozestan, Iran, in 1985. He received the Ph.D. degrees in Electrical Engineering from Shahid Rajaee Teacher Training University in 2019. From 2019 to 2022, he was an Assistant Professor of Electrical Engineering /Communications Engineering at the Department of Electrical Engineering Azad University, Iran. His research interested in (Antenna, Analog and Digital circuit design, RF and Microwave circuits). From June 2022 to now he has been working as ADAS Technology at ZF Group, Germany.

- Email: a.kiani@sru.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Fatemeh Geran was born in Ghaemshar, Iran in 1977. She received her B.Sc. degree in Electrical Engineering (Telecommunication) from Tehran University, Tehran, Iran in 1999. Also, she received her M.Sc. and Ph.D. degrees in Electrical Engineering (Telecommunication) from Tarbiat Modares University, Tehran, Iran, in 2003 and 2009, respectively. She is currently an Associate Professor in the Faculty of Electrical Engineering at Shahid Rajaee Teacher Training University, Tehran, Iran. Her research interest fields are antenna, RF subsystems in the microwave and mm-wave bands, and RF energy harvesting.

- Email: f.geran@sru.ac.ir
- ORCID: 0000-0002-7845-8391
- Web of Science Researcher ID: AAN-8757-2020
- Scopus Author ID: 16038985700
- Homepage: https://www.sru.ac.ir/geran-2/

Seyed Mohammad Hashemi was born in Tehran, Iran, in 1983. He received the B.Sc., M.Sc. and Ph.D. degrees in Electrical Engineering from the Iran University of Science and Technology (IUST), Tehran, Iran, in 2006, 2008 and 2013, respectively. From 2012 to 2013 he joined Aalto University, Finland, as a Visiting Scholar. Since 2015, he is an Assistant Professor of Communications Engineering at the Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University (Tehran Iran). His research interests include Applied Electromagnetics, Optimization Methods, Antenna and Microwave Engineering.

- Email: sm.hashemi@sru.ac.ir
- ORCID: 0000-0003-1484-9008
- Web of Science Researcher ID: X-8115-2019
- Scopus Author ID: 57192714769
- Homepage: https://www.sru.ac.ir/hashemi/

How to cite this paper:
A. Kiani, F. Geran, S. M. Hashemi, "A new approach to synthesis of a quasi non-uniform leaky wave antenna," J. Electr. Comput. Eng. Innovations, 12(1): 227-234, 2024.

DOI: 10.22061/jecei.2023.9919.662

URL: https://jecei.sru.ac.ir/article_1998.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A New High-Speed Multi-Layer Three-Bits Counter Design in Quantum-Dot Cellular Automata Technology

*G. Asadi Ghiasvand, M. Zare *, M. Mahdavi*

Department of Electronics, Faculty of Engineering, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran.

Article Info

Article History:
Received 24 July 2023
Reviewed 28 September 2023
Revised 29 October 2023
Accepted 05 November 2023

Keywords:
Digital circuit design
Quantum-Dot-Cellular Automata
Multi-layer design
Three bits counter
T flip-flop

*Corresponding Author's Email Address:
d.mehdi.zare@gmail.com

Abstract

Background and Objectives: Quantum-dot Cellular Automata technology is a new method for digital circuits and systems designs. This method can be attractive for researchers due to its special features such as power consumption, high calculation speed and small dimensions.
Methods: This paper tries to design a three-bits counter with minimum area and delay among the other circuits. As the circuit dimensions are reduced, the area and consequently, the delay are decreased, too. Therefore, this paper tries to design a three-bits counter with minimum dimensions and delay. The proposed counter contains 96 cells and is designed in three layers. It has the least area and delay compared to the priors.
Results: The circuit simulation illustrates 0.08 μm^2 of area occupation and one clock cycle delay. In comparison with the best previous design, which includes 110 cells, the cells number, area and delay are decreased by 12.72%, 27.27% and 33.33%, respectively. Also, the cost of the circuit has been improved by 54.32%. The power analysis of the design shows 13% reduction in the total energy dissipation of the circuit compared to the best prior work. The circuit reliability versus temperature variations has been simulated and the results represent suitable stability. The fault tolerance of the circuit which is occurred by the displacement faults represents normal operation of the circuit.
Conclusion: As the counter is an element which is implemented in several digital systems, its area reduction causes the whole system area to be reduced. Also, the circuit delay has been decreased significantly which means that the circuit can be employed by high speed systems.

Introduction

With the advancement of technology and moving towards nanotechnology, circuit speed and dimensions are becoming the important factors that need to be improved. One of the proposed technologies besides the complementary metal oxide semiconductor (CMOS) [1] is the quantum dot cellular automata (QCA), which is a novel nanoscale design method proposed by Lent in the 1993. The QCA is a nano-dimension technology and the digital circuits can be designed and implemented by this method [2]-[6].

One of these circuits is the counter which is widely used in digital systems. In [7], a counter has been presented by JK flip-flop, which has 278 cells, area of 0.33 μm^2 and delay of 2 clock cycles. The cells number and area of this circuit are high. In [8], a counter by three consecutive T flip-flops has been proposed, which has 244 cells, area of 0.346 μm^2 and delay of 4.25 clock cycles. The design advantage is a NOT gate, which prevents signal weakness. In [9], a 3-bits counter has been designed by T flip-flop, which has 238 cells, area of 0.361 μm^2 and delay of 2.25 clock cycles. In this design,

the circuit delay and cost have been improved compared to the circuit in [8].

In [10], a D flip-flop counter has been presented by 196 cells which has good delay and cost. A counter in [11], has employed D flip-flop and has 174 cells, area of 0.194 μm^2 and delay of 3 clock cycles. This design has improved cells number and area, but is not optimal in terms of delay and cost. In [12], a counter with 140 cells has been presented by T flip-flop and each T flip-flop contains one XOR logic gate. A counter with 110 cells has been proposed in [13] that is the best design in terms of delay, area and cost compared to the priors. As shown in Fig. 1, the main advantage of the circuit in [13] is the crossing wires at two different clock phases. So, the wires cross each other with the minimum area and delay.

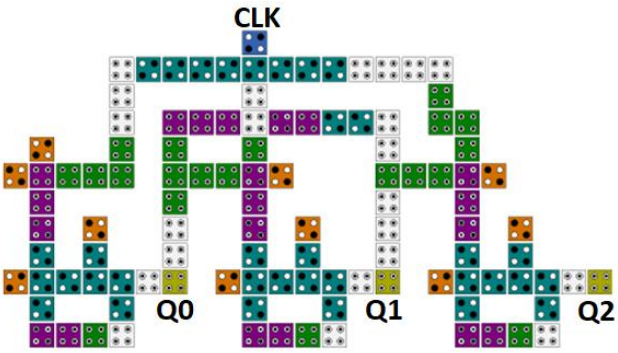


Fig. 1: Structure of three-bits counter with 110 cells [13].

Main contributions: We design a new 3-bits counter in this paper and follow these contributions:

1- It has been tried to design a circuit with the minimum number of cells.
2- Since the counter delay is related to the delay of T flip-flops, the implemented T flip-flops structures have been redesigned to improve the circuit delay. This will qualify the circuit for high-speed applications.
3- The total energy dissipation of the circuit has been improved.
4- It has been tried to preserve the circuit stability and the circuit fault tolerance in normal situation.

In continuous of the paper, the background of the QCA is discussed in section 2. In Section 3, the circuit design method is introduced and the proposed multi-layer 3-bits counter is explained. Simulation results and comparisons, the circuit energy analysis, and reliability are represented in section 4 and finally, the conclusion is given in section 5.

QCA Background

The QCA is a nano-electronics computing architecture and is implemented to build circuits in nano-dimensions [14]. The basic unit in the QCA is its cell. As shown in Fig. 2, the QCA cell shape is square and consists of four quantum dots [15]-[16].

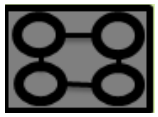
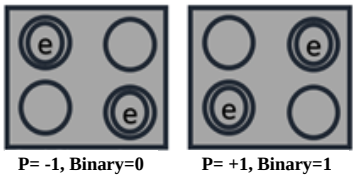
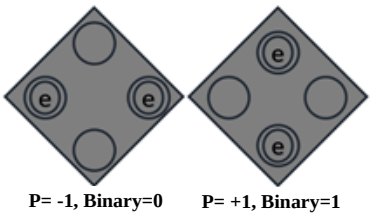


Fig. 2: Structure of a QCA cell.

The cell with four quantum dots, positioned at the corners of a square, contains two free electrons. An electron can be quantum-mechanically confined in a quantum dot. Due to the Coulombic interaction, electrons occupy the farthest possible positions from each other [17]. Therefore, the two stable states are the corners of the QCA cell and consequently, the possible polarizations are P = "-1" and P = "+1", which are translated as logics '0' and '1', respectively. The logics of the QCA cell are shown in Fig. 3(a) and 3(b). The standard distance between two adjacent cells is 2 nm and the cell standard dimensions are 18 x 18 nm^2, as shown in Fig. 4 [10].



(a)



(b)

Fig. 3: (a) The angle of electron is 90. (b) The angle of electron is 45.

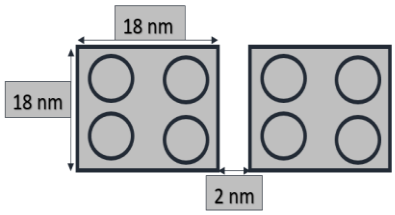
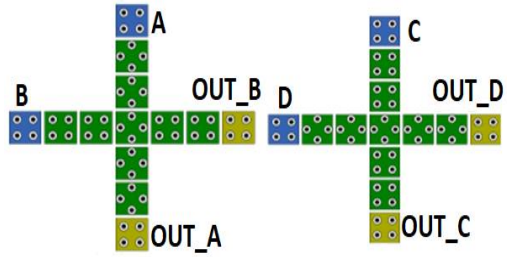


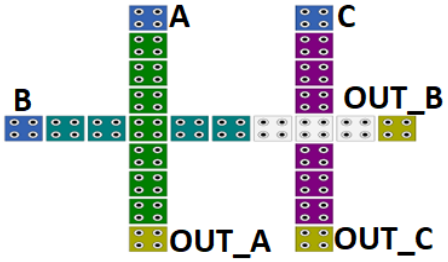
Fig. 4: Quantum dot cell dimensions.

The Coulombic interaction effects the adjacent cells and transfers the logic of one cell to the other cells. Therefore, the consecutive cells can create a QCA wire [18]. Since there is no current and output capacitor in the circuit, the QCA has lower power consumption compared to the CMOS. The QCA has simple structures for 'AND' and 'OR' gates, and has the ability to cross the

wires over each other. In classical binary QCA, wires crossing can be implemented by two methods that are coplanar wires and multi-layer wires. In the coplanar, the cells can cross with two different orientations or with two different clock phases as shown in Fig. 5. In multiple layers, one of the wires is transferred to the second layer and then returns to the first layer. Thus, two wires cross without affecting each other. Fig. 6 shows the overview of the multi-layer wires. It is proved that the multi-layer approach is more robust than the coplanar method [19]. In coplanar crossing, the wire coupler is loose and can be affected by the random external influences. In multi-layer, intermediate layers are used to prevent any possible crosstalk. Each layer can be employed as an active layer on which a new circuit can be designed independently [20]. A multi-layer model decreases the overall area and thus potentially consumes less chip area compared to the coplanar approach.



(a)



(b)

Fig. 5: Coplanar wires (a) Two wires crossing in 45& 90-degrees. (b) Two wires crossing with 2 different clock phases.

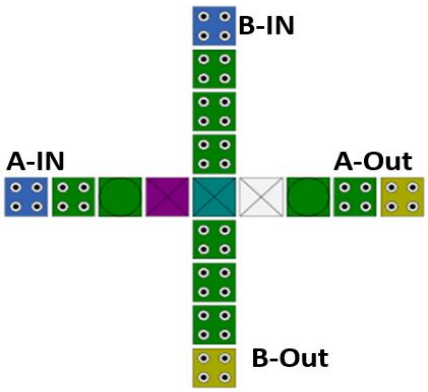
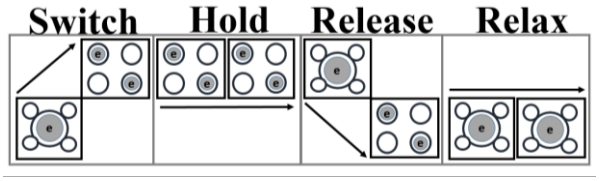
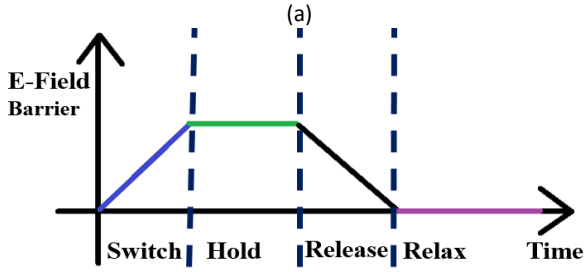


Fig. 6: Multi-layer wires crossing.

Binary logic movement in the QCA requires clocking. Timing/synchronization in the QCA is done by the cascaded clocking of four distinct and periodic phases as shown in Fig. 7.



(a)



(b)

Fig. 7: (a) Clock diagrams in QCA. (b) Clock phases in QCA.

The phases are rising (switch), falling (Release), staying at high potential (Hold) and staying at low potential (Relax). The applications of clocking are:

1- Creating an appropriate mechanism for synchronous movement of data in the circuit.

2- Determining the direction of data movement in the circuit.

3- Supplying the required power for the operation of the circuit.

A tool for the QCA circuits simulations is “QCA Designer”. In the “QCA designer” tool, the phase of the cells, along with the type of the cells (input/output), are determined by a special coloring which are shown in Fig. 8.

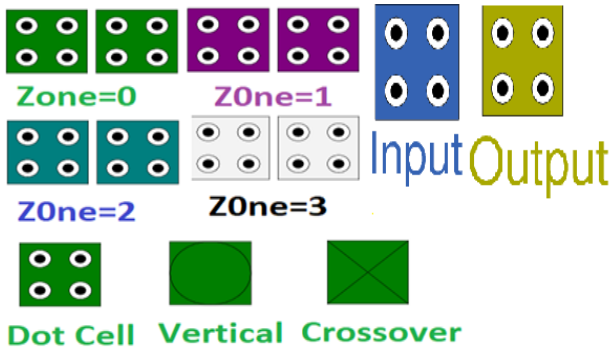


Fig. 8: The clock phases of the QCA and coloring of the cells.

As mentioned, by arranging the QCA cells in a line, the polarity of the first cell is transferred to the last cell and acts like a wire. In general, two wiring methods are utilized in the QCA. The first method is standard wiring and is accomplished as shown in Fig. 9(a). The second

method in Fig. 9(b) is called complementary chain, which cells are rotated 45 degrees with "+1" and "-1" polarization. The advantage of this type of wiring is its ability to transfer both the input signal and the reverse of the input signal in odd and even cells, respectively.

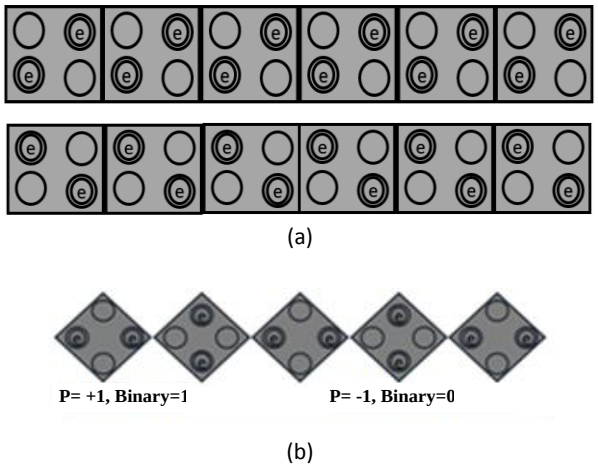


Fig. 9: (a) standard wiring. (b) complimentary chain wiring.

Proposed Circuit

The interference minimization between wires and clock synchronization can be implemented in multi-layer designs [21]-[22]. Also, the multi-layer method reduces the overall area of the chip compared to the coplanar designs [23]. Therefore, to design the circuit with minimum area, the proposed circuit has been implemented with 3-layers. It has three output bits and contains 96 cells.

A. Design Methodology

To propose a new circuit, we have employed the circuit in [13] as the base circuit and follow this algorithm to improve it:

1- Try to optimize the elements of the circuit. This will help you to optimize the whole circuit parameters.

2- Try to minimize the circuit dimensions. If the circuit dimensions are reduced, you have the opportunity to reduce the circuit delay.

3- If at least one of the previous steps has been performed successfully, simulate the circuit to see whether the circuit parameters are improved or not. If the parameters are improved, go to the next step, otherwise the circuit is not a good candidate as the new circuit.

4- Check the stability of the circuit. If it is stable, the new circuit has been designed correctly, otherwise the proposed circuit must be redesigned again in the first or second steps.

The flowchart of the circuit design methodology is depicted in Fig. 10.

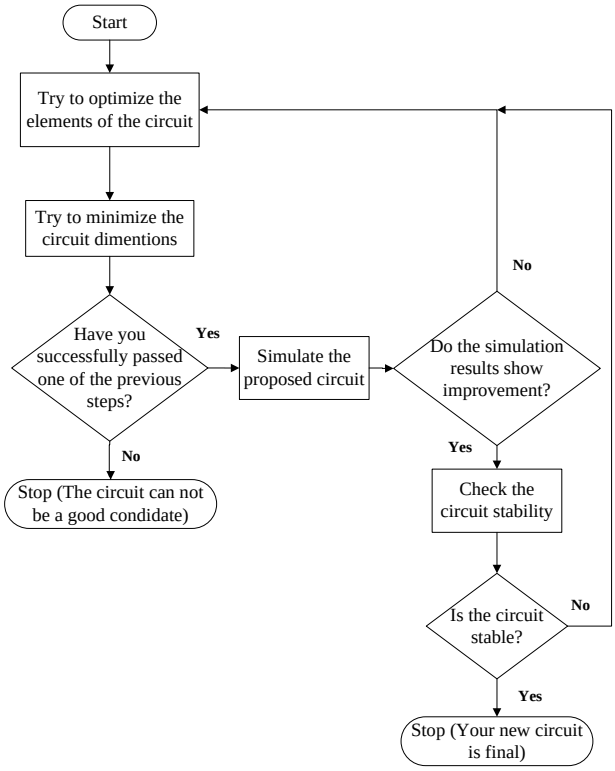


Fig. 10: The flowchart of the design methodology.

B. T Flip-Flop Structure

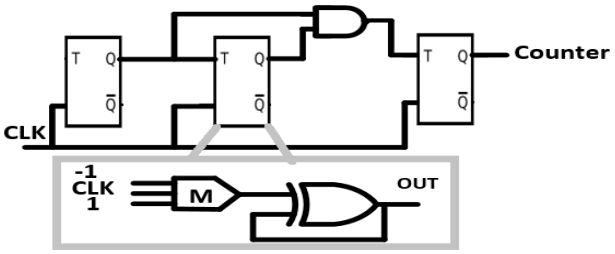
A counter is constructed by the flip-flops. According to the flowchart, we try to improve the flip-flops of the circuit in this section. The block diagram of the circuit in Fig. 11(a) shows three consecutive T flip-flops. The structure of each T flip-flop includes one 'XOR' gate and one 'AND' gate. The 'AND' gate is constructed by a three inputs majority gate which its inputs are "+1" and "-1" constant polarizations and the CLK signal. Totally, six majority gates have been utilized to design 'AND' and 'XOR' gates. Fig. 11(b) shows the internal structures of 'AND' and 'XOR' gates.

The block diagram of the T flip-flop is similar to [13]. However, as shown in Fig.12, its structure has the following differences:

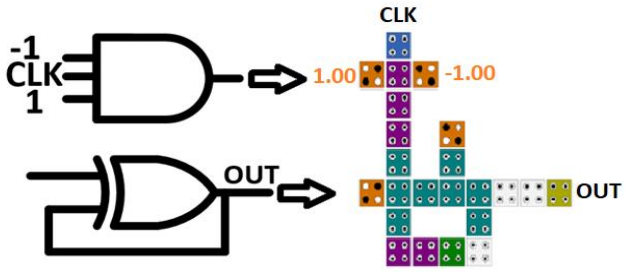
1- To reduce the number of cells in the 'XOR' majority gates of the first and third T flip-flops, normal cells employ phase 1 and fixed cells employ phase 2 of the clock. This will synchronize the data current and controls the electrons movement in the cells.

2- The input clock is applied directly to each T flip-flop which improves the circuit delay.

These simple modifications lead to the first aim of our research and decrease the delay of the T flip-flops. Consequently, by modification of the T flip flops, the total circuit delay is reduced and causes the circuit to be a candidate for high speed applications.

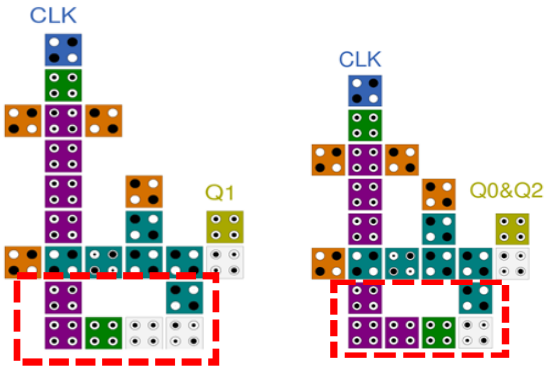


(a)



(b)

Fig. 11: (a) Block diagram of three bits counter. (b) Structure of 'AND' and 'XOR' gates.



(a) (b)

Fig. 12: (a) the structure of second flip-flop. (b) the structure of first and third flip-flop.

C. Circuit Design

The second step in the flowchart is the minimization of the circuit dimensions. To achieve this goal, the following modifications have been performed: 1- the wires routings have been changed. 2- three layers are selected for the circuit. 3- the vertical wire of the first and third T flip-flops has been diminished by one cell. The schematic of the circuit is illustrated in Fig. 13. As shown, one of the wires is taken to another layer and then returns to the original layer. Thus, two wires cross without affecting each other. In brief, the proposed circuit has the following modifications compared to the circuit in [13]:

1- We have replaced the crossover in prior work with the multi-layer crossover.

2- The circuit of the T flip-flop is redesigned and optimized as described in Fig. 12.

3- The wires routings have been changed to decrease the number of vertical and horizontal cells of the counter.

4- The outputs of 'AND' gates are connected directly to the 'XOR' gates which reduce one cell in each XOR gate.

These modifications lead to the area reduction aim of our research, so that the circuit cells number is reduced to 96 cells. Also, by the modifications in the T flip-flop circuit, the delay is reduced to one clock cycle.

Simulation Results and Discussion

We have implemented "QCA Designer" tool to design the proposed counter. This tool has two simulation engines, which are "Coherence Vector" and "Bistable Approximation". In this research, the Coherence Vector engine is utilized and its parameters are depicted in Fig. 14.

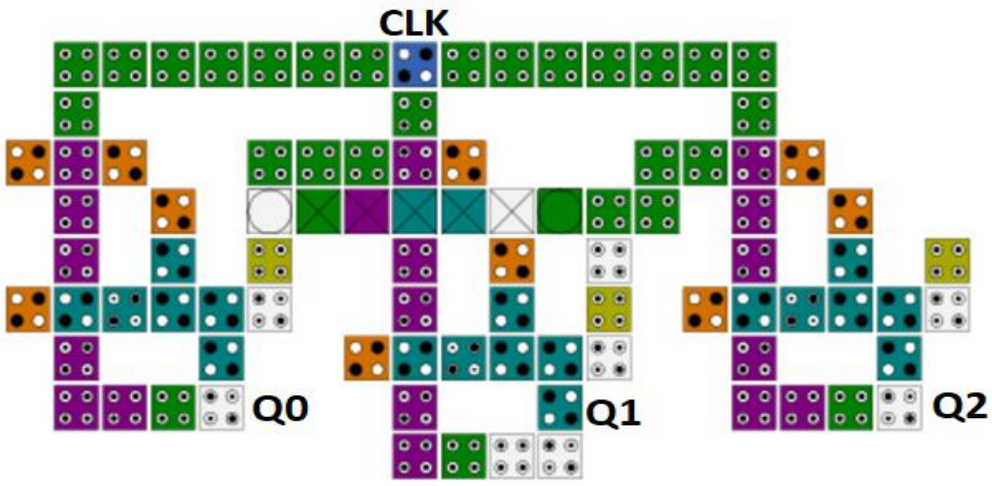


Fig. 13: The structure of proposed multilayer 3-bits counter.

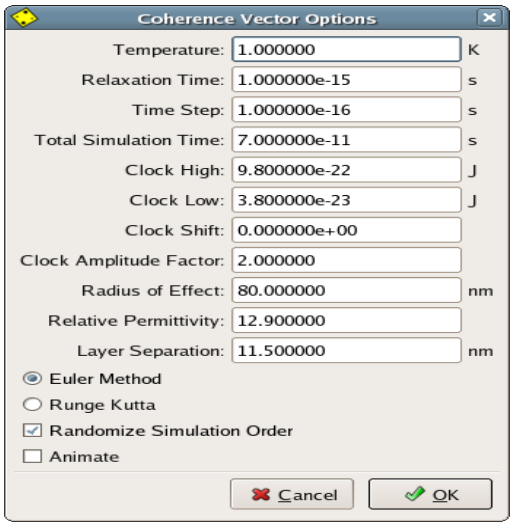


Fig. 14: The simulation parameters of coherence vector engine in QCA Designer tool.

The simulation waveforms of the proposed counter are represented in Fig. 15.

As seen, the numbers are counted properly and the circuit works correctly.

Fig. 16 shows the cells number of the proposed design in comparison with the previous designs. As depicted, the circuit contains 96 cells which is lowest among the others. The area of our circuit and the prior circuits are drawn in Fig. 17.

The occupied area of the proposed design is 0.08 μm^2, which is the least occupied area.

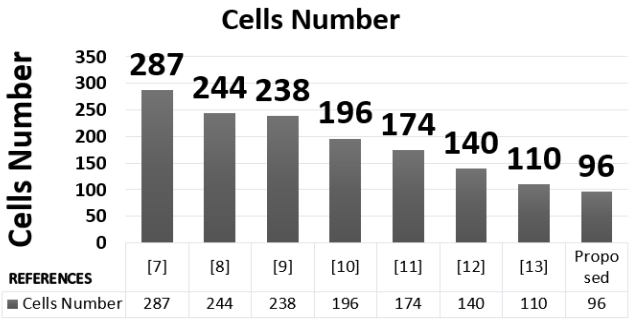


Fig. 16: The comparison of the cells number.

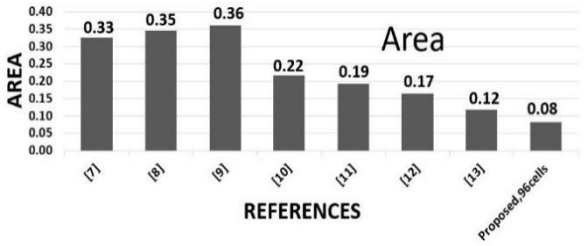


Fig. 17: Area occupied of the proposed design and previous studies.

The circuit delay is calculated by the number of clock cycles that the input signal takes time to reach to its output. The delay of our circuit is depicted in Fig. 18. As seen by the colors of the cells, the path from 'CLK' signal to the 'Q2' has experienced four consecutive phases of the clock which is equal to one clock cycle. As expected, the simulation of the proposed circuit shows that the delay is 1 clock cycle. The curve in Fig. 19 represents the delays of various designs.

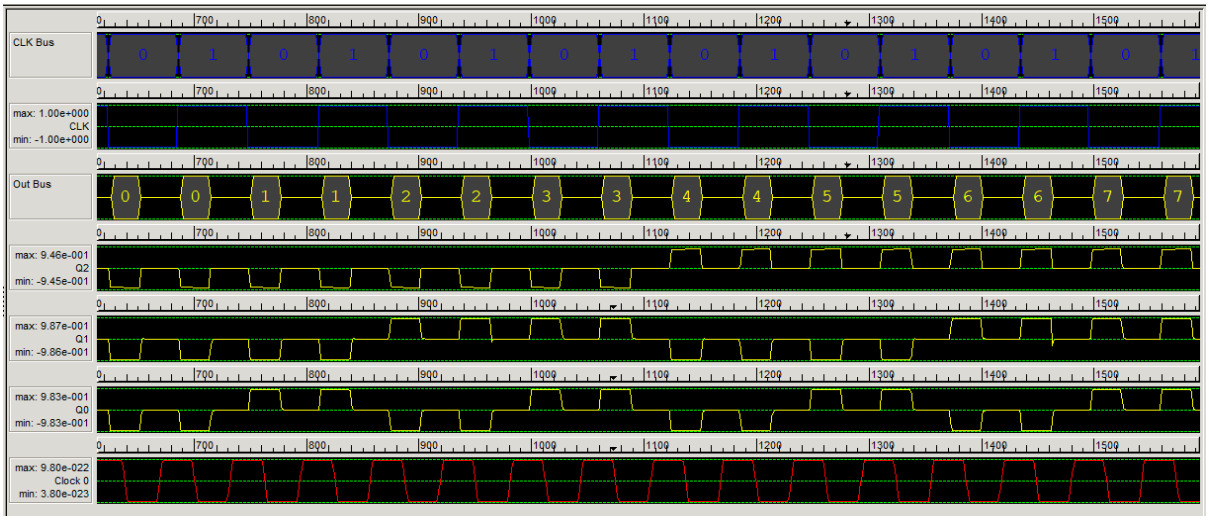


Fig. 15: Simulation results of 3-bit multilayer counter with 96 cells.

Table 1: Simulation results and comparison with the priors

parameters	[7]	[8]	[9]	[10]	[11]	[12]	[13]	Proposed 96 Cells	our circuit improvement compare to [13]
Cell Count	287	244	238	196	174	140	110	96	12.72%
Cell/Bit	95.67	81.33	79.33	65.33	58	46.67	36	32	11.11%
layers number	1	1	1	1	1	1	1	3	----
Length Covered	778	958	858	638	438	458	458	418	8.73%
Width Covered	418	358	418	338	438	358	238	198	16.81%
Net Area	92988	79056	77112	63504	56376	45360	35640	31104	12.73%
Total Area	325204	342946	358644	215644	191844	165600	109004	82764	24.07%
Area	0.33	0.35	0.36	0.22	0.19	0.17	0.11	0.08	27.27%
Area/Bit	0.11	0.12	0.12	0.07	0.06	0.06	0.04	0.03	25%
Wasted Area	819	816	903	544	484	414	276	210	23.91%
Latency or delay	2	4.25	2.25	2	3	2	1.5	1	33.33%
Cost	348	885	855	688	324	196	81	37	54.32%

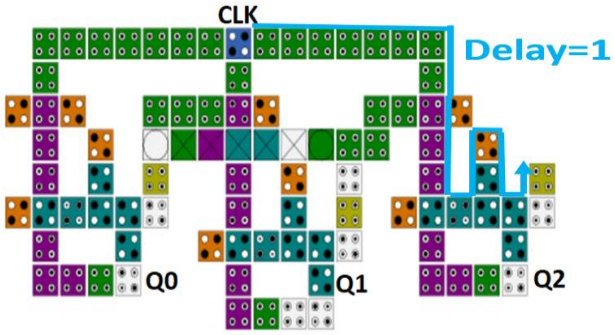


Fig. 18: The path from 'CLK' to the output 'Q2'.

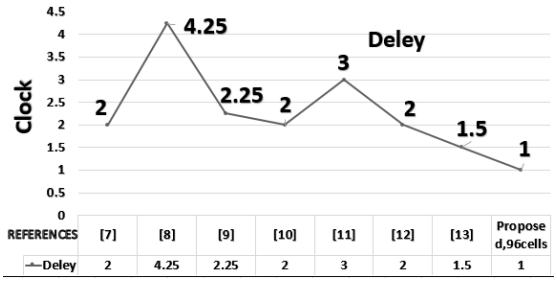


Fig. 19: The comparison of delay.

Table 1 shows all the parameters of the proposed design and the previous studies. The work in [13] has better parameters compared to the others, and therefore, our circuit has been compared to it. As seen, the delay has improved by 33.33%. The cells number decrement is 12.72% and the area reduction is 27.27%. A QCA specific cost function estimation is proposed in [24] that takes into account the number of logic gates and crossings wires in cost evaluation.

Since our circuit is multi-layer, we implement this function for the evaluation of cost metric:

$$Cost = (M^2 + I + C^2) \times T^2 \quad (1)$$

where (T) shows the circuit delay, (M) represent the number of majority gates, (I) is the number of inverters and (C) is the number if crossovers. As the number of layers and consequently, the number of crossovers increases, we expect the cost to be incremented. However, as there is only one crossover in the circuit and the delay is decreased to one clock cycle, the circuit cost has been reduced by 54.32%.

The circuit Net area is calculated by multiplication of the cells number and the area of one cell. The Net area is decreased by 12.73%.

Net area = cells number × area of one cell (2)

The circuit total area is calculated as represented in (3). It is seen that the total area is reduced by 24.07%.

Total Area = Width Covered×Length Covered (3)

Width Covered=(Vertical cells number×20nm)–2 (4)

Length Covered=(horizontal cells number×20nm)–2 (5)

To calculate the energy consumption of the circuit, "QCA Designer-E" tool is employed. This tool analyses "average energy dissipation" and "total energy dissipation" and represents the results in an output log text. The energies comparison of the proposed circuit and the priors are illustrated in Table 2. As seen, both the total energy dissipation and the average energy dissipation per clock cycle have been improved by 13%.

Table 2: The results of the energy dissipation in the proposed circuit

Reference	Total energy dissipation	Average energy dissipation	our circuit improvement
[7]	8.66E-02	7.87E-03	68%
[8]	7.53E-02	6.84E-03	63%
[9]	6.63E-02	6.03E-03	58%
[10]	5.87E-02	5.34E-03	53%
[11]	3.39E-02	3.08E-03	18%
[12]	3.73E-02	3.39E-03	26%
[13]	3.18E-02	2.89E-03	13%
Proposed circuit	2.75E-02	2.50E-03	----------

A. Circuit Stability and Reliability

Temperature Stability Factor (TSF) is a parameter that calculates the reliability and accuracy of the circuit. The TSF indicates the temperature range in which the circuit operates correctly. The temperature variations can disrupt the circuit operation and make it unstable. The TSF is determined by the minimum and maximum polarizations in the circuit simulation [25], [26]. The TSF simulation is performed by changing "Temperature option (k)" parameter of the coherence vector engine in the "QCA Designer" tool. Table 3 illustrates the TSF analysis of the proposed circuit. As seen, it is stable from k=1 to k=6 which is similar to the TSF of [13]. The TSF of [12] is from k=1 to k=4 which is lower than our design. Fig. 20 represents the TSF(Min) and TSF(Max) of the proposed circuit. It shows that after k=6, all curves are diverged from each other and the circuit operates incorrectly.

B. Fault Tolerance Analysis

Displacement faults are usual in the QCA cell production. Therefore, it needs to be attended in the QCA circuit design [27], [28]. The Defects can occur in the individual cells production or by the movement of cells to a surface [29]-[30]. To evaluate the fault tolerance, each cell should move separately in 5 directions which are vertical (up and down), horizontal (right and left) and 45-degrees rotation. Also, the cell missing is considered in the simulation. The segmented design of the counter is illustrated in Fig. 21. Tables 4 & 5 show the fault tolerance simulation for all cells of the circuit. Some cells such as C13R4 and C15R3 are critical and their movement to horizontal right direction make the circuit unstable. Also, there are a number of cells such as C6R6 that their movement range in all directions are wide.

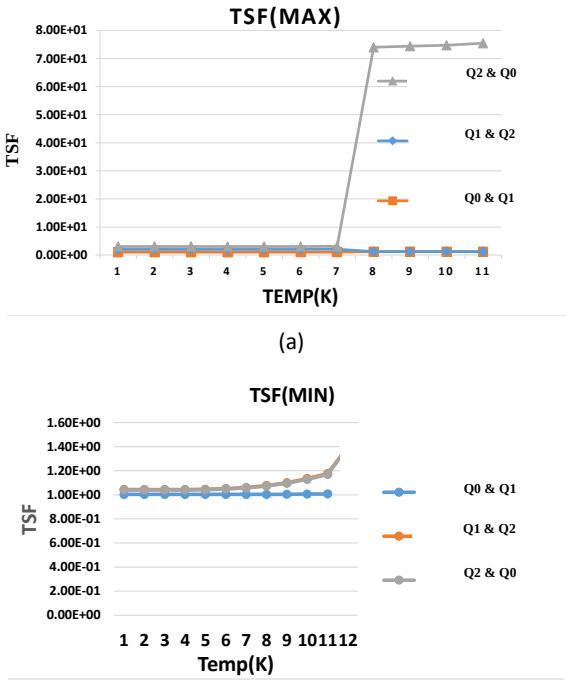


Fig. 20: Temperature Stability Factor (a) TSF(Max) (b) TSF(Min).

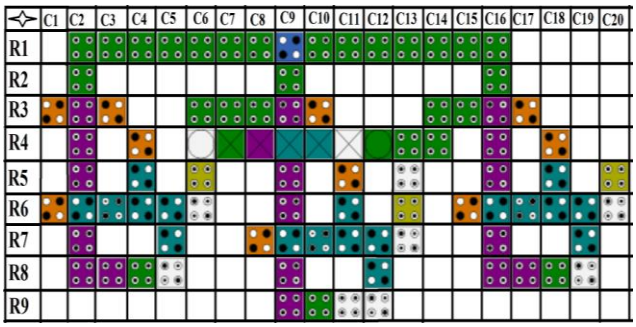


Fig. 21: Segmented design of the three bits counter.

Table 3: Temperature stability factor for output of the counter

Temp (k)	TEMPERATURE STABILITY (MIN & MAX) FOR-three-bits counter with 96 cells						state
	Q0&Q1		Q1&Q2		Q2&Q0		
	TSF (MAX)	TSF (MIN)	TSF (MAX)	TSF (MIN)	TSF (MAX)	TSF (MIN)	
1	1.00	1.00	1.04	1.04	1.04	1.04	**stable**
2	1.00	1.00	1.04	1.04	1.04	1.04	**stable**
3	1.00	1.00	1.04	1.04	1.04	1.04	**stable**
4	1.00	1.00	1.04	1.04	1.04	1.04	**stable**
5	1.00	1.00	1.05	1.05	1.04	1.04	**stable**
6	1.00	1.00	1.05	1.05	1.05	1.05	**stable**
7	1.00	1.00	1.06	1.06	1.05	1.06	**unstable**
8	0.0169	1.00	1.23	1.08	72.7	1.07	**unstable**
9	0.0168	1.00	1.23	1.10	73.1	1.10	**unstable**
10	0.0168	1.01	1.23	1.14	73.5	1.13	**unstable**

For the majority of cells, the rotation by 45 degree and cell missing make the circuit unstable. The exceptions are for a few numbers of cells such as C1R3, C3R3 and etc. Generally, we can not conclude exactly that the output is fluctuation free, because a few numbers of cells are critical and their movement can make the circuit unstable.

On the opposite side, there are several cells that can be displaced more than 1.5nm. So, it depends deeply to the fabrication process accuracy and its fault percentage. However, from the calculated values in Tables 4 & 5, we can summarize that about 70% of the cells are allowed to be displaced by 1.5nm or more.

Table 4: Fault tolerance analysis (C1 to C9)

Name of Cells	Horiz. Left (nm)	Horiz. Right (nm)	Vertic. Up (nm)	Vertic. Down (nm)	Rotate (45 DEG)	Missing Cell
C1R3	stable	2.5	7.8	5.4	stable	stable
C1R6	0.5	2.2	2.2	2.2	unstable	unstable
C2R1	5.7	2.5	5.5	3.6	unstable	unstable
C2R2	5.1	4.3	3.6	4.7	unstable	unstable
C2R3	1.7	1.4	4.8	6.8	unstable	unstable
C2R4	4.6	4.1	2.8	5.1	unstable	unstable
C2R5	4.2	1.5	1.6	0.6	unstable	unstable
C2R6	0.3	0.6	0.2	0.4	unstable	unstable
C2R7	3.8	1	0.6	0.9	unstable	unstable
C2R8	2.6	3.5	3.8	1.1	unstable	unstable
C3R1	2.5	3.4	stable	5.6	stable	stable
C3R3	2.3	34	7.7	7.2	stable	stable
C3R6	3	0.7	1	0.4	unstable	unstable
C3R8	2.8	5.1	3.6	3.9	unstable	unstable
C4R1	2.5	5.6	stable	stable	unstable	unstable
C4R4	3.8	3.8	1.6	0.4	unstable	unstable
C4R5	2.3	1.7	1.2	0.2	unstable	unstable
C4R6	0.8	0.1	0.1	0.5	unstable	unstable
C4R8	1.1	5.8	2.7	2.8	unstable	unstable
C5R1	5.5	6.6	5.6	5.5	unstable	unstable
C5R6	0.6	3.8	1.9	1.5	unstable	unstable
C5R7	3.7	2.5	2.8	4.1	unstable	unstable
C5R8	6.2	0.9	3.9	1.2	unstable	unstable
C6R1	6	6.8	5.4	5.6	unstable	unstable
C6R3	1.6	6.1	1.7	4.4	unstable	unstable
C6R4-L1	3.8	2.7	4.4	2.2	unstable	unstable
C6R4-L2	6.8	6.7	6.6	5.1	unstable	unstable
C6R4-L3	stable	3.9	stable	stable	stable	stable
C6R5	1.8	4.1	1.6	6.6	unstable	unstable
C6R6	5	stable	5.8	stable	unstable	unstable
C7R1	6.7	7.2	6.2	6.3	unstable	unstable
C7R3	6.6	1.3	5.8	3.7	unstable	unstable
C7R4-L3	3	6.1	3.5	4.4	unstable	unstable
C8R1	6.8	14.6	6.9	6.3	unstable	unstable
C8R3	3.6	2.8	5.4	3.3	unstable	unstable
C8R4-L3	5.8	6.7	3	2.8	unstable	unstable
C8R7	0.6	2.2	2.4	2.4	unstable	unstable
C9R1	5.1	5.1	3.4	5.8	unstable	unstable
C9R2	3.7	3.8	8.4	1.9	unstable	unstable
C9R3	1.7	1.3	1	5.6	unstable	unstable
C9R4-L1	3.5	5.2	7	2.2	unstable	unstable
C9R4-L3	7.1	6.8	4.4	4.5	unstable	unstable
C9R5	4.5	4.6	2.7	2.2	unstable	unstable
C9R6	4.2	1.2	1.2	0.9	unstable	unstable
C9R7	0.4	0.6	0.2	0.3	unstable	unstable
C9R8	4	1.1	0.5	1	unstable	unstable
C9R9	1.5	4	3.9	1	unstable	unstable

Table 5: Fault tolerance analysis (C10 to C20)

Name of Cells	Horiz. Left (nm)	Horiz. Right (nm)	Vertic. Up (nm)	Vertic. Down (nm)	Rotate (45 DEG)	Missing Cell
C10R1	9.5	6.8	4.8	4.5	unstable	unstable
C10R3	2.2	3.5	5.3	3.3	unstable	unstable
C10R4-L3	6.7	6.1	3.5	3.4	unstable	unstable
C10R7	3.4	0.7	0.7	0.5	unstable	unstable
C10R9	6.8	5	2.7	2.6	unstable	unstable
C11R1	6.2	6.7	3.5	3.5	unstable	unstable
C11R4-L3	0.4	6.8	1.9	1.6	unstable	unstable
C11R5	3.8	0.8	1.1	0.5	unstable	unstable
C11R6	2.2	1.7	1.1	0.2	unstable	unstable
C11R7	0.9	0.2	0.1	0.5	unstable	unstable
C11R9	5.3	6.5	3.8	3.8	unstable	unstable
C12R1	6.8	6.6	2.1	4.1	unstable	unstable
C12R4-L1	1.5	0.1	3.6	2.8	unstable	unstable
C12R4-L2	1	2.5	1.8	2.6	unstable	unstable
C12R4-L3	6.2	0.2	1.4	0.9	unstable	unstable
C12R7	0.7	3.4	3.5	3.1	unstable	unstable
C12R8	6.9	5.7	2.2	4.4	stable	stable
C12R9	6.3	6	4.1	5.9	unstable	unstable
C13R1	6	0.6	1.1	4.4	unstable	unstable
C13R4	0.1	0	0.2	0	unstable	unstable
C13R5	0.1	0.8	1.6	3.7	unstable	unstable
C13R6	3.7	1.5	2.5	1.3	unstable	unstable
C13R7	4.8	2.9	5.9	2.4	unstable	unstable
C14R1	6.6	1	4.4	0.6	unstable	unstable
C14R3	0.05	0.1	0.1	0	unstable	unstable
C14R4	0	0.1	0	0	unstable	unstable
C15R1	0.3	1.9	4.7	0.7	unstable	unstable
C15R3	0.2	0	0.5	0.2	unstable	unstable
C15R6	0.5	2.2	2.2	2.2	unstable	unstable
C16R1	2.3	0.4	0.1	0.9	unstable	unstable
C16R2	1.5	0.1	1.3	0.1	unstable	unstable
C16R3	1	0.3	0.1	1.8	unstable	unstable
C16R4	1.6	2.4	3	2.7	unstable	unstable
C16R5	4.1	1.4	1.6	0.8	unstable	unstable
C16R6	0.3	0.6	0.2	0.4	unstable	unstable
C16R7	3.8	1	0.5	0.9	unstable	unstable
C16R8	2.6	3.5	3.8	1.1	unstable	unstable
C17R3	0.3	0	1.1	0.4	unstable	unstable
C17R6	2.9	0.7	0.9	0.4	unstable	unstable
C17R8	6.8	5	3.6	3.8	unstable	unstable
C18R4	5.2	5.4	3.8	0.4	stable	unstable
C18R5	4	4	4	0.2	unstable	unstable
C18R6	0.8	0.1	0.1	0.5	unstable	unstable
C18R8	1.2	5.8	2.7	2.8	unstable	unstable
C19R6	0.6	3.8	4	1.3	unstable	unstable
C19R7	3.8	2.6	2.9	4	unstable	unstable
C19R8	6.2	0.9	3.9	1.3	unstable	unstable
C20R5	4.3	5.7	4.5	6.6	unstable	unstable
C20R6	5.1	4.7	6	3.1	unstable	unstable

Conclusion

A three-bits QCA counter utilizing T flip-flop with 96 cells was introduced in this research. The circuit designed in three layers and its outputs are three bits. The design methodology was introduced and our design flowchart was explained. The counter components were drawn in a block diagram and each component explained with details. The whole circuit along with its routing and clocking methods were illustrated and the modifications were defined. The modifications contain two parts which were T flip-flops improvement and circuit dimensions reduction. These modifications were performed to minimize the circuit delay and to make the circuit as a high-speed counter. The simulation waveforms of the circuit proved the design accuracy. The simulation results depicted that the proposed circuit has been improved in delay, area and energy dissipations compared to the previous designs. So that the terms of cells number, area and delay were improved by 12.72%, 27.27%, and 33.33%, respectively compared to [13]. Also, the total energy dissipation and the average energy dissipation showed 13% improvement. The circuit temperature stability and fault tolerance of the circuit was analyzed. The temperature analysis illustrated that the circuit temperature stability was similar to the previous circuit and it was not decreased. The fault tolerance simulation represented that about 70% of the cells can be displaced by 1.5 nm or more. We can say somehow that the fault tolerance is dependent to the fabrication process and it is a weakness for the proposed circuit.

Author Contributions

G. Asadi-Ghiasvand designed the experiments. G. Asadi-Ghiasvand, M. Zare and M. Mahdavi collected and carried out the data analysis. G.Asadi-Ghiasvand and M. Zare interpreted the results and wrote the manuscript.

Acknowledgment

The author would like to thank the Shahr-e-Qods Branch, Islamic Azad University, that sponsored this research.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

CMOS	Complementary Metal Oxide Semiconductor
QCA	Quantum-dot Cellular Automata
P	Polarization
CLK	Clock
T	Circuit delay
M	Number of majority gates
I	Number of inverters
C	Number if crossovers
TSF	Temperature Stability Factor
k	Temperature parameter in QCA-Designer

References

[1] K. Jeong, A. B. Kahng, “A power-constrained MPU roadmap for the International Technology Roadmap for Semiconductors (ITRS),” in Proc. 2009 international SoC Design Conference (ISOCC), 2009.

[2] R. Chakrabarty, D. Kumar Mahato, A. Banerjee, S. Choudhuri, M. Dey, N. K. Mandal, "A novel design of flip-flop circuits using quantum dot cellular automata (QCA)," in Proc. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC): 408-414, 2018.

[3] H. Cho, E. E. Swartzlander, “Adder and multiplier design in quantum-dot cellular automata,” IEEE Trans. Comput., 58(6): 721-727, 2009.

[4] M. A. Dehkordi, A. S. Shamsabadi, B. S. Ghahfarokhi, A. Vafaei, “Novel RAM cell designs based on inherent capabilities of quantum-dot cellular automata,” Microelectron. J., 42(5): 701-708, 2011.

[5] M. N. Asfestani, S. R. Heikalabad, “A novel multiplexer-based structure for random access memory cell in quantum-dot cellular automata,” Phys. B: Condens. Matter, 521: 162-167, 2017.

[6] C. S. Lent, P. D. Tougaw, W. Porod, G. H. Bernstein, “Quantum cellular automata,” Nanotechnology, 4(1): 49-57, 1993.

[7] S. Angizi, S. Sayedsalehi, A. Roohi, N. Bagherzadeh, K. Navi, "Design and verification of new n-bit quantum-dot synchronous counters using majority function-based JK flip-flops," J. Circuits Syst. Comput., 24(10): 1550153, 2015.

[8] K. S. Bhavani, V. Alinvinisha, "Utilization of QCA based T flip flop to design Counters," in Proc. 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS): 1-6, 2015.

[9] S. Angizi, M. H. Moaiyeri, S. Farrokhi, K. Navi, N. Bagherzadeh, "Designing quantum-dot cellular automata counters with energy consumption analysis," Microprocess. Microsyst., 39(7): 512-520, 2015.

[10] M. Abutaleb, “Robust and efficient quantum-dot cellular automata synchronous counters,” Microelectron. J., 61: 6-14, March. 2017.

[11] Z. Amirzadeh, M. Gholami, “Counters designs with minimum number of cells and area in the quantum-dot cellular automata technology,” Int. J. Theor. Phys., 58: 1758-1775, 2019.

[12] A. H. Majeed, E. Alkaldy, M. S. bin Zainal, D. Bin MD Nor, “Synchronous counter design using novel level sensitive T-FF in QCA technology,” J. Low Power Electron. Appl., 9(3): 27, 2019.

[13] J. Mohammadi, M. Zare, M. Molaei, M. Maadani, "Low-cost three-bit counter design in quantum-dot cellular automata technology," IETE J. Res., 2022.

[14] C. S. Lent, P. D. Tougaw, “Lines of interacting quantum-dot cells: A binary wire,” J. Appl. Phys., 74(10): 6227-6233, 1993.

[15] F. Lombardi, J. Huang, "Design and test of digital circuits by quantum-dot cellular automata," Artech House, Inc., Electronic ISBN:9781596932685, 2007.

[16] P. D. Tougaw, C. S. Lent, "Logical devices implemented using quantum cellular automata," J. Appl. Phys., 75(3): 1818-1825, 1994.

[17] C. S. Lent, P. D. Tougaw, "A device architecture for computing with quantum dots," Proc. IEEE, 85(4): 541-557, 1997.

[18] B. Sen, M. Dalui, B. K. Sikdar, "Introducing universal qca logic gate for synthesizing symmetric functions with minimum wire crossings," in Proc. ICWET: 828–833, 2010.

[19] G. Schulhof, Konrad, G. A. Jullien, "Simulation of random cell displacements in qca," J. Emerg. Technol. Comput. Syst., 3(1), 2007.

[20] K. Walus, G. Schulhof, G. A. Jullien, "High level exploration of Quantum-Dot Cellular Automata (QCA)," in Proc. 38th Asilomar Conference on Signals, Systems and Computers: 30–33, 2004.

[21] A. Gin, P. D. Tougaw, S. Williams, "An alternative geometry for quantum-dot cellular automata," J. Appl. Phys., 85(12): 8281–8286, 1999.

[22] I. L. Bajec, P. Pecar, "Two-layer synchronized ternary quantum-dot cellular automata wire crossings," Nanotechnol., 7: 368–376, 2012.

[23] B. Sen, A. Nag, A. De, B. K. Sikdar, "Multilayer Design of QCA Multiplexercellular," in Proc. 2013 Annual IEEE India Conference (INDICON), 2013.

[24] W. Liu, L. Liang, M. O'Neill, E. Swartzlander, "A first step towards cost functions for quantum-dot cellular automata designs," IEEE Trans. Nanotechnol., 13(3): 476-487, 2014.

[25] S. S. Roy, "Fault tolerance and temperature stability: the dynamic error estimation in quantum-dot cellular automata," in proc. 2017 IEEE International Symposium on Nanoelectronic and Information Systems (iNIS): 84-89, 2017.

[26] D. Bhowmik, J. Pal, M. Chandra, A. K. Saha, N. Kumar, "QCA based design of cost-efficient code converter with temperature stability and energy efficiency analysis," Mater. Today Proc., 49(8): 3585-3594 , 2022.

[27] B. Sen, A. Nag, A. De, B. K. Sikdar, "Towards the hierarchical design of multilayer QCA logic circuit," J. Comput. Sci., 11: 233-244, 2005.

[28] J. Pal, S. Bhattacharjee, A. K. Saha, P. Dutta , " Study on temperature stability and fault tolerance of adder in quantum-dot cellular automata," in Proc. 5th IEEE International Conference on Signal Processing, Computing and Control (ISPCC 2k19), 2019.

[29] M. Momenzadeh, H. Jing, M. B. Tahoori, F. Lombardi, "Characterization, test, and logic synthesis of and-or-inverter (AOI) gate design for QCA implementation," IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 24 (2005): 1881-1893, 2005.

[30] A. Roohi, R. F. DeMara, N. Khoshavi, "Design and evaluation of an ultra-area-efficient fault-tolerant QCA full adder," Microelectron. J., 46(6): 531-542, 2015.

Biographies

Ghodratollah Asadi Ghiasvand received his B.Sc. in Telecommunication Engineering, transmission trend from Ghiaseddin Jamshid Kashani university, Abyek, Iran in 2011 and his M.Sc. in Telecommunication systems trend from Islamic Azad university، Department of Electrical Engineering, Shahr-e-Qods Branch, Tehran, Iran, 2022. He has been working in the radio department of Infrastructure Communications Company since 2012, He is currently an expert in the data department of Infrastructure Communications Company and interests in digital circuit design and telecommunication system design.

- Email: telecom.engineer.asadi@gmail.com
- ORCID: 0009-0003-7959-4496
- Web of Science Researcher ID: JLM-1309-2023
- Scopus Author ID: NA
- Homepage: NA

Mahdi Zare received his B.Sc. and M.Sc. degrees in Electrical Engineering from South Tehran Branch, Islamic Azad University, Tehran, Iran, in 2001 and 2004, respectively. He received his Ph.D. degree in Electronic Engineering at the Tehran Science and Research Branch, Islamic Azad University. He worked as senior hardware engineer in Rayaphone Co. since 2004. He is currently the faculty member of Shahr-e-Qods branch, Islamic Azad University. His research interests include performance and area optimization in latency-insensitive systems, multi-core synchronization, and Mixed signals circuit design. He has published several scientific papers and has acted as a reviewer in several international journals and conferences.

- Email: zare@qodsiau.ac.ir, d.mehdi.zare@gmail.com
- ORCID: 0000-0002-9797-0083
- Web of Science Researcher ID: AAO-1254-2021
- Scopus Author ID: 57129066700
- Homepage: NA

Mojdeh Mahdavi received the Ph.D. degree in Electronics from Islamic Azad University, Science and Research Branch, Tehran, Iran, in 2010. In 2006, she joined the Electrical Engineering Department as an Assistant Professor at Islamic Azad University, Shahr-e-Qods Branch, Tehran, Iran. Her current research interests include digital system design and implementation, nanotechnology and fault tolerant design.

- Email: mahdavi.qodsiau@gmail.com
- ORCID: 0000-0001-8774-2426
- Web of Science Researcher ID: NA
- Scopus Author ID: AAO-1261-2021
- Homepage: NA

How to cite this paper:
G. Asadi Ghiasvand, M. Zare, M. Mahdavi, "A new high-speed multi-layer three-bits counter design in quantum-dot cellular automata technology" J. Electr. Comput. Eng. Innovations, 12(1): 235-246, 2024.

DOI: 10.22061/jecei.2023.9955.667

URL: https://jecei.sru.ac.ir/article_1999.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Motif-Based Community Detection: A Probabilistic Model Based on Repeating Patterns

H. Hajibabaei [1], V. Seydi [1,2,*], A. Koochari [2]

[1]*Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.*

[2]*Centre for Applied Marine Sciences, School of Ocean Sciences, Bangor University, Menai Bridge, UK.*

Article Info

Article History:
Received 13 August 2023
Reviewed 15 October 2023
Revised 05 November 2023
Accepted 04 December 2023

Keywords:
Community detection
Motif
Complex networks
Probabilistic model

*Corresponding Author's Email Address: *V.seydi@bangor.ac.uk*

Abstract

Background and Objectives: The detection of community in networks is an important tool for revealing hidden data in network analysis. One of the signs that the community exists in the network is the neighborhood density between nodes. Also, the existence of a concept called a motif indicates that a community with a high edge density has a correlation between nodes that goes beyond their close neighbors. Motifs are repetitive edge patterns that are frequently seen in the network.

Methods: By estimating the triangular motif in the network, our proposed probabilistic motif-based community detection model (PMCD) helps to find the communities in the network. The idea of the proposed model is network analysis based on structural density between nodes and detecting communities by estimating motifs using probabilistic methods.

Results: The suggested model's output is the strength of each node's affiliation to the communities and detecting overlaps in communities. To evaluate the performance and accuracy of the proposed method, experiments are done on real-world and synthetic networks. The findings show that, compared to other algorithms, the proposed method is acting more accurately and densely in detecting communities.

Conclusion: The advantage of PMCD in using the probabilistic generative model is speeding up the computation of the hidden parameters and establishing the community based on the likelihood of triangular motifs. In fact, the proposed method proves there is a probabilistic correlation between the observation of two node pairs in different communities and the increased existence of motif structure in the network.

Introduction

For effective network component monitoring and recognition, network analysis is a key tool. A complex network [1] can consist of cells in biology [2], social networks with friendly communication [3], or a network of scientists doing joint scientific studies [4]. To put it another way, it can be any grid with nodes and edges that can be represented as a graph. One of the most effective methods and methodologies to analyze complex networks is community detection.

Community detection identifies subgraphs of a network whose relationship among their nodes is more robust and dense than those between other nodes of the network [5]. A community can represent an idea, a group, an interest, a focus on a particular topic, and so on. Communities can be used separately or together; the latter are referred to as overlap communities. The machine learning clustering topic's component,

community detection, has the potential to be applied in a variety of engineering fields, such as text classification, traffic network optimization, and social network analysis. The goal of community detection is to group the nodes of a network into different communities so that they are strongly connected or have similar node features [5]. A key problem in dynamical network research is the discovery of communities with the aim of revealing hidden features of a complex network, which are frequently densely coupled nodes [6].

Community detection in networks is an NP-hard issue that categorized from different perspectives. These categories include weighted [7], [8] and unweighted [9], [10], directed [11] and undirected [12], global [13] and local [12], overlapping [14], [15], and non-overlapping [16] community discovery techniques. Different community detection techniques were developed based on these criteria. Examples include model-based approaches [7], [14], [17], clique percolation methods [18], modularity-based methods [4], [19], [20], label propagation methods [11], [21]-[23], model-based methods [7], [14], methods for network embedding [24] and community detection methods with deep learning [25]-[27]. It can be seen from examining the different approaches used for community detection and the research on this topic that a straightforward analysis of node properties won't produce the accuracy needed for community detection in networks; rather, taking a deeper look at the networks' particulars and using the graphs' original characteristics, like motif structure, will produce better results.

The PMCD method employs a probabilistic relation to find communities in complex networks. We extend probabilistic model-based methods from edge creation to motif generation. Complex networks commonly contain "motifs", which are a type of small, linked sub-networks. Based on empirical studies, communities with similar nodes have related motifs. As a result, using motifs with lots of connections can be a useful strategy for finding communities and performing more accurate network analysis [28]. We show that the chance of a triangle motif existing between three nodes in shared communities grows with the observation of more nodes in such communities. In other words, we locate the hidden parameter of the probabilistic model and find the community by using the triangle motif. We define the triangular motif estimator function as a Bernoulli loss function over one node and two of that node's neighbors for the probabilistic motif generator's function. We also research how community overlap affects how motifs are generated.

Related Works

The problem of community detection in complex networks gets a lot of attention. Several research projects on different aspects of community detection have been performed over the past few years. The first methods of community detection employed traditional techniques and clustering-based algorithms. These methods presented key ideas for community detection and laid the groundwork for future developments. Traditional approaches include graph partitioning, hierarchical, spectral and partitional clustering [29].

The algorithms used to detect communities based on modularity have been extensively studied and used due to their simple tactics and clear outcomes. However, they also encounter difficulties, such as communities that are unstable and sensitive to seed node selection [30]. One of these, the Louvain technique [5], is frequently applied to weighted graphs. This approach provides a straightforward and quick methodology to detect distinct communities and maximizes modularity by clustering network nodes using the greedy approaches [31]. The Leiden method, however, corrects numerous flaws in the Louvain algorithm [32]. The objective is to change the community developed throughout the iteration cycle while simultaneously speeding up local mobility and transferring nodes to arbitrary neighbors.

The label propagation algorithm (LPA), a practical community detection approach, was initially introduced in [22]. Although its simple design and low complexity are widely respected, there are several downsides, such as the randomness of node selection and label updating. In the LPA technique, a node is selected at random, and through an iterative process, its label is updated with the most prevalent label nearby [33]. To handle the weaknesses in the LPA methodology, the Speaker-Listener Label Propagation Algorithm (SLPA) [34] and the COPRA [35] were created.

Cliques are one of the fundamental ideas in graph analysis and are utilized to detect communities in networks. The clique percolation algorithm (CPM) [18] and CFinder [36] were proposed as overlapping community detection algorithms based on the clique percolation method's search for local patterns.

The motif is another idea related to the clique. Small, linked sub-networks known as motifs frequently appear in complex networks and are one of the basic elements of the network [37]. In network analysis, motifs are used to detect communities and comprehend network structure [38]. Motifs demonstrate that a community with a high edge density will have relationships between nodes that go beyond their immediate neighbors. Although a few motif-based community detection methods have been proposed [38]-[40], when used on large-scale networks, they frequently encounter high computational complexity. It is still challenging to properly and economically combine lower-order and higher-order structural data into a unified framework for community detection.

The group of methods estimates the probabilistic model to detect communities, in difference to the techniques cited at the top that employ traditional methods to do so. This method creates a generative

sample of the graph and estimates the model parameters [14], [17]. The degree of node dependency on communities is a parameter in the generative model that is estimated using methods [14], [17], [41] using a matrix factorization-based model. The algorithm [42] presents a matrix factorization-based paradigm that makes it easy to add or delete edges. The non-negative matrix factorization model of the community detection issue is also described [43] and a transfer matrix is then used to control the dynamics of the network structure.

Proposed Algorithm Frameworks

In this research, a probabilistic motif-based community detection model (PMCD) is presented that uses the triangle motif and the affiliation graph model to detect community structures. The core idea of the proposed community detection method is that a robust community requires taking the node's structural model and relationship types into consideration. Two nodes observed in more shared communities are more likely to be connected, according to Yang and Leskovec's study [14] on the connection between edge (2-clique) likelihood and community overlap. In this paper, we examine the impact of community overlapping on the evolution of the triangle and 3-clique motifs. We show that by increasing the number of nodes observed in shared communities, the probability of the existence of a triangular motif between them increases. This result is in accordance with the fundamental principle that vertices situated in communities' overlaps are more densely connected than vertices within a single community. By using the optimized Bernoulli loss function for probabilistic estimation, we can therefore enhance and increase the AGM's ability [14], [17] to generate triangle motifs.

The PMCD model is different from other community detection approaches in that it considers additional properties that were not sufficiently considered in earlier methods, for instance:

- Using edge density to detect communities.
- Triangle motif estimation using a probabilistic approach.
- The conceptual link between community detection and the likelihood of the triangular motif being present or absent.
- Use of evolutionary approaches and maximum likelihood estimation in computations.

Fig. 1 presents an example of a simple graph to illustrate the concept of three-node motifs. Two different three-node motif types discovered in Fig. 1 are shown in Fig. 2. Depending on the characteristics of the networks, the triangle motifs observed in various types of networks can be interpreted in multiple ways. For instance, the 3-node motif (3, e) and the 4-node motif (4, e) are the most widely studied motifs in complex networks [44]. The proposed method uses the triangular motif to build the hidden parameter of the probabilistic model and detect the community.

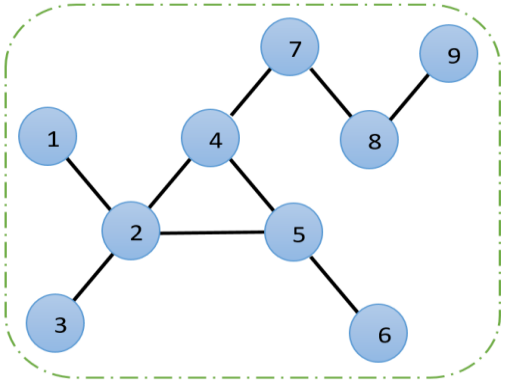


Fig. 1: Illustrated an example of a simple graph to illustrate the concept of three-node motifs.

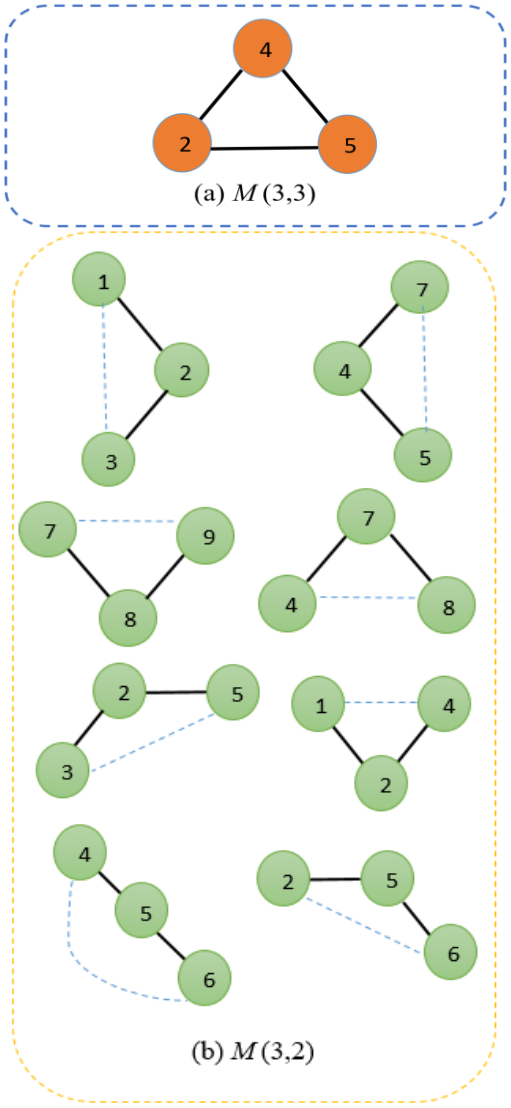


Fig. 2: Illustrated two types of three-node motifs that we use in the proposed model: (a) a 3-clique or closed triangle motif (denoted as M(3,3)-motif) with 3 nodes and 3 edges discovered from Fig. 1; (b) types of the opened triangle motif (denoted as M(3,2)-motif) with 3 nodes and 2 edges extracted from Fig. 1.

The PMCD model is based on a network G (N, E), where nodes and edges are referred to as N and E, respectively. We create M_{uc}, a nonnegative integer, to represent the strength of the node's affiliation with the community. (M_{uc} = 0 denotes u's non-membership in c.) The degree of reliance between each node and each community is thus shown in the M matrix.

The value of M in PMCD establishes whether or not a triangular motif between three nodes (u, v_1, and v_2) will appear in a community (c). Specifically, we presumed that three nodes, u, v_1, and v_2, are triangular motifs by taking into account the following likelihood. For the probabilistic motif generator's function, we define the triangular motif estimator function as a loss function over one node and two neighbors of that node, that is,

$$P_c(u,v_1,v_2) = \underset{u,v_1\in E}{P_c}(u,v_1).\underset{u,v_2\in E}{P_c}(u,v_2) = \left[1-\exp(-M_{uc}\cdot M_{v_1c}^T)\right]\cdot\left[1-\exp(-M_{uc}\cdot M_{v_2c}^T)\right] \quad (1)$$

Due to the generative probabilistic approach between two couples of nodes in a triangle motif, each couple of nodes is independently propagated by the Bernoulli model. Thus, per element of the adjacency matrix is formed on the following probabilistic method:

$$P(u,v_1,v_2) = \underset{v_1\in N(u)}{P}(u,v_1).\underset{v_2\in N(u)}{P}(u,v_2) = \left[1-\exp(-M_u\cdot M_{v_1}^T)\right]\cdot\left[1-\exp(-M_u\cdot M_{v_2}^T)\right] \quad (2)$$

$$A_{uv_1} \sim Bernoulli(P_{uv_1}) \odot A_{uv_2} \sim Bernoulli(P_{uv_2})$$

The framework of computation for (1) and (2), which describe a probabilistic generative model, is predicated on the following premises:

- In a community, a triangle motif can exist between two pairs of nodes (one node and two neighbors of that node).
- The probability of the existence triangle motif increases when two pairs of nodes are observed in multiple communities.
- Communities can overlap; communities that overlap have a higher density of triangle motifs.

Community detection by PMCD model

We defined the components of the PMCD model before illustrating how to utilize it for community detection in networks. The model parameter discussed in the preceding section is the degree of a node's community membership (M_{uc}). By maximizing the likelihood, we can get the optimum M as follows:

$$likelihood(M) = l(M) = log\ P(G\mid M)$$

$$\hat{M} = \arg\max l(M) = \arg\max_{M\geq 0} \prod_{(u,v)\in v} p(u,v_1,v_2) \prod_{(u,v)\notin v}(1-p(u,v_1,v_2)) = \arg\max_M \left[\prod_{(u,v)\in v} p(u,v_1)\cdot p(u,v_2)\right]\cdot \left[\prod_{(u,v)\notin v}(1-(p(u,v_1)\cdot p(u,v_2))\right] \quad (3)$$

After combining (2) and (3), a natural logarithm is needed to be computed on both sides to correct the multiplication to the aggregate and reduce the next computations. The logarithm is completely ascending; therefore, it won't interfere with the maximum likelihood estimation.

$$L(M) = \left[\sum_{(u,v_1)\in E}\log(1-\exp(-M_u.M_{v_1}^T)) - \sum_{(u,v_1)\notin E} M_u.M_{v_1}^T\right] + \left[\sum_{(u,v_2)\in E}\log(1-\exp(-M_u.M_{v_2}^T)) - \sum_{(u,v_2)\notin E} M_u.M_{v_2}^T\right] \quad (4)$$

Updating the Parameter

The non-linear likelihood function of (4), which contains the latent variable M, cannot be maximized by conventional optimization methods. We calculate the objective function in (4) using the *Block Coordinate Ascent* approach [45], which helps solve optimization problems with latent variables in machine learning. By maintaining fixed neighbors (M_v), we update M_u for each node u.

$$L(M_u) = \left[\sum_{v\in N(u)}\log(1-\exp(-M_u.M_{v_1}^T)) - \sum_{v\notin N(u)} M_u.M_{v_1}^T\right] + \left[\sum_{v\in N(u)}\log(1-\exp(-M_u.M_{v_2}^T)) - \sum_{v\notin N(u)} M_u.M_{v_2}^T\right] \quad (5)$$

In (5), *N(u)* is a set of neighbours of u. In order to calculate the maximum probability (the diagram's maximum point), we must find a location on the Figurative chart where the gradient equals 0. Thus, it is necessary to derive the partial derivation of the likelihood logarithmic in (5) than M_u.

$$\frac{\partial\ell(M_u)}{\partial M_u} = \left[\sum_{v_1\in N(u)} M_u \frac{\exp(-M_u M_{v_1})}{1-\exp(-M_u M_{v_1})} - \sum_{v_1\notin N(u)} M_{v_1}\right] + \left[\sum_{v_2\in N(u)} M_u \frac{\exp(-M_u M_{v_2})}{1-\exp(-M_u M_{v_2})} - \sum_{v_2\notin N(u)} M_{v_2}\right] \quad (6)$$

The gradient ascent algorithm will eventually update M_u values [46], [47]. A node's belonging strength to a community will be replaced with 0 if it detects it, as it is impossible for it to be negative.

$$M_u(t+1) = \max\left(0, M_u(t) - \eta\ \left(\frac{\partial\ell(M_u)}{\partial M_u}\right)\right) \quad (7)$$

In (5), η is a learning parameter, As long as the difference between the value from the last step and the current value is lower than the acceptable threshold, the process of updating each M_u at each stage of the algorithm iteration is repeated.

PMCD Algorithm

Algorithm 1 displays the proposed PMCD model (probabilistic motif-based community detection). A graph

(G) and the number of communities (k) are the method's inputs. The model also creates a matrix (M_{uc}) that displays the degree to which each node belongs to each community. When they are observed in different communities, the likelihood that there is an existing motif structure between two groups of nodes increases.

Since the hidden variable (M) is initialized (details of computing M addressed later), the method then begins an iterative process. After the difference among $M_u(t+1)$ and $M_u(t)$ was smaller than a predefined point (in this case, the stop threshold is 0.005), the iterations stop. In order to estimate the model's unknown parameter in the graph, this iterative method calculates the likelihood of the probabilistic model ($L(M_u)$). To derive the likelihood function's logarithm as close as possible to its maximum value (when the line's slope is 0), the likelihood function's logarithm is collected from each node u using the formula $D(L(M_u))$.

Algorithm 1: Probabilistic motif-based community detection (PMCD)

1: Inputs: Graph G = (N; E);
Number of communities (k);
2: Output: M_{uc} belonging of each node *u* Community c
3: $t \leftarrow 0$
4: M = *local_maximun_neighborhood()*
5: while $|M_u(t+1) - M_u(t)| <= 0.005$ do
6: t ← t + 1
7: **for** i = 1 to |V| do
8: $L(M) = log\ p(G|M)$
9: $D(L(M_u))$ = *Derivation_finder_L*(M_u)
10: Update: $M_u\ (t+1) =$
Gradient_ascent $(D(L(M_u)); M_u(t))$
11: **end for**
12: end while
13: for i = 1 to |V| do
14: **for** j = 1 to k do
15: **if** $M_{uc} > threshold$ then
16: Add: $c_j \leftarrow u_i$
17: **end if**
18: **end for**
19: end for

In line 10 of Algorithm 1**,** we chose the ascending gradient approach [46], [47] to optimize the probability because the computations were complex. This method is used to update the latent variable of the model (Mu) at each iteration of the algorithm. After the M value has been fixed, each node's ability to contribute per community is assessed. Since comparing this value to a testing point (such as the median of M), it may be defined as either belonging to or not belonging to the communities, and the output of the model will then be realized.

Computational Complexity

The count of communities and dense motifs affect the computational complexity of the PMCD method. The core concept of Algorithm 1, as shown in its iteration phases, is the rate of depending on the community, which is updated using (6) and (7). In this case, whether or not two nodes have neighbors who are members of one or more communities determines whether or not those nodes share a theme. Because of this, the computational complexity will depend on the number of communities present and the order of each node's neighbors (N(u)); in the worst case, this complexity will be O(2k.|E|).

Initialize

The matrix of depending strengths for the nodes communities can start in a variety of ways. The first option, which also appears to be the simplest, involves filling in the values at random. The algorithm's major drawback, however, is that it repeats the steps more frequently, increasing computing complexity as it advances to the model stability phases.

The other choice is the local minim neighborhood approach [48], which has been shown through studies to be an excellent starting point for community discovery algorithms. Using this method has the added benefit of being able to estimate the initial number of communities to start the proposed model's community detection phase, in addition to minimizing iteration steps and starting the process in a stable state.

Experiments

The proposed PMCD method has been implemented in the Spyder environment using the Python programming language. We used five real-world data sets (Table 2) and sixteen synthetic networks (Table 5), respectively, to evaluate the results. Additionally, the statistics include the node's "ground-truth" community memberships. In these datasets, the proposed method is compared with fundamental algorithms like Louvain [49], Leiden [32], Bigclam [14], [17], CPM [18], Label propagation [35], and SLPA [34]. Table 1 lists these algorithms in brief.

Evaluation Metrics

We evaluate the community detection algorithms' effectiveness and accuracy using three standard evaluation metrics. Modularity [50] is an internal metric for assessing community quality, whereas the F1_Score and NMI are external metrics for assessing community accuracy by comparing them to ground-truth communities [6]. The modularity measure in internal metrics, a popular benchmark for estimating the density in the community is derived from Girvan-Newman [50].

Table 1: The employed methods for PMCD evaluation

Method Name	Description
Louvain	Louvain amplifies the modularity value of communities
Leiden	The Leiden method is an advancement of the Louvain
Bigclam	The probabilistic community detection method that scales to large networks
CPM	Find k-clique communities in a graph using the percolation method
LPA	The label propagation algorithm detects communities by network structure
SLPA	SLPA is an overlapping community discovery that extends the LPA

By dividing the projected community edges by the expected community edges, the modularity value is calculated. The identified community will perform better if there are more nodes inside the community and if the modularity score of the community is around 1. When comparing the frequency of properly recognizing the nodes in each community using the supplied ground truth data, the F1_Score is a well-known evaluation statistic used in community detection methods. The other outsider statistic is NMI, or mutual information, about the connection found among the recognized groups and the real world.

Real-World Datasets

Five real-world datasets are used in the experiments. Zachary's karate club network [51] is the first dataset, containing 34 nodes, 78 connecting edges between them, and 2 ground-truth communities. This dataset contains social ties among university karate club members collected by Wayne Zachary in 1977. Dolphins' online social network [52] is the second dataset, which contains 62 nodes, 159 connecting edges, and two ground-truth communities containing a list of all the links, where a link represents frequent associations between dolphins. The third dataset [53], with 105 nodes, 441 connecting edges, and 3 ground-truth communities, is based on data from the network of books about US politics published around the time of the 2004 presidential election. Edges between books represent frequent co-purchasing of books by the same buyers. The fourth dataset is the American football [4], with 116 nodes, 613 connecting edges, and 12 ground-truth communities. This network contains American football games between Division IA colleges during the fall of 2000. The fifth dataset is a large network generated using email data from a large European research institution [54], [55]. This network contains 1005 members of the institution as nodes, and 25571 edges contain emails sent between members of the institution and people outside of the institution. The dataset also assumes departments at the research institute as the nodes' ground-truth community memberships. Each individual belongs to exactly one of the 42 departments at the research institute.

The real-world datasets analyzed during the current study are shown in Table 2, where N is the number of nodes, E is the number of edges, and K is the number of ground truths. These datasets are available in the network repository[1] [56], the KONECT project[2] [53], and the Stanford Network Analysis Project[3] [54] (SNAP).

Table 2: The specifics of the real-world dataset used

Dataset Name	N (#Nodes)	E (#Edges)	K (#Ground_truth)
Karate	34	78	2
Dolphin	62	159	2
Pol-Book	105	441	3
Football	115	613	12
Email-EU	1005	25571	42

[1] https://networkrepository.com/
[2] http://konect.cc/
[3] https://snap.stanford.edu/

Experimental on Real Datasets

We evaluate the PMCD by four kinds of community detection models, such as modularity optimization, label propagation, probabilistic estimation, and clique percolation, in order to assessment the efficacy and accuracy of PMCD in community detection. In the sections before, several of these methods were briefly discussed. The suggested approach is assessed using six algorithms using internal evaluation criteria (modularity and community number) as well as external evaluation metrics (NMI and F1_Score).

The findings in Table 3 demonstrate that our method has more accuracy than other methods in terms of the internal metrics (modularity maximum and accuracy in the number of communities).

Additionally, Fig. 3 and 4 demonstrate that our suggested method has absolute superiority over probabilistic estimation and clique percolation methods and relative superiority over modularity optimization and label propagation methods in the external assessment criteria (NMI and F1_Score).

Table 3: Experimental results on real-world networks by the modularity metric (Q) and community number (CN)

Methods	Louvain		Leiden		Bigclam		CPM		LPA		SLPA		PMCD	
Datasets	Q	CN	Q	CN	Q	CN	Q	CN	Q	CN	Q	CN	Q	CN
Karate	0.415	4	0.116	5	0.204	4	0.215	4	0.354	3	0.371	3	0.397	3
Dolphin	0.518	5	0.134	7	0.185	6	0.321	5	0.456	4	0.470	3	0.522	3
Pol-Book	0.526	4	0.279	9	0.347	8	0.271	9	0.481	5	0.493	5	0.543	4
Football	0.604	10	0.257	19	0.381	16	0.283	18	0.552	14	0.596	13	0.632	11
Email-EU	0.432	27	0.226	58	0.207	65	0.162	74	0.274	55	0.303	52	0.507	46

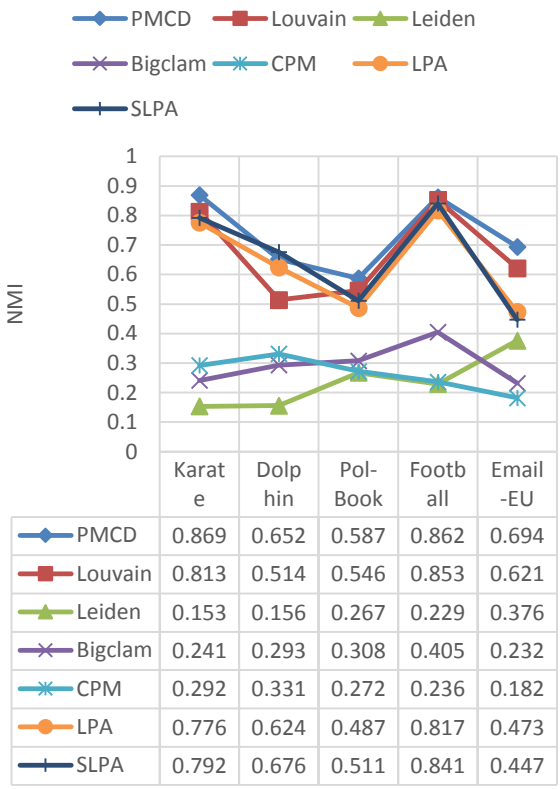


Fig. 3: NMI assessment chart, compare PMCD by community detection models on five real-world data sets.

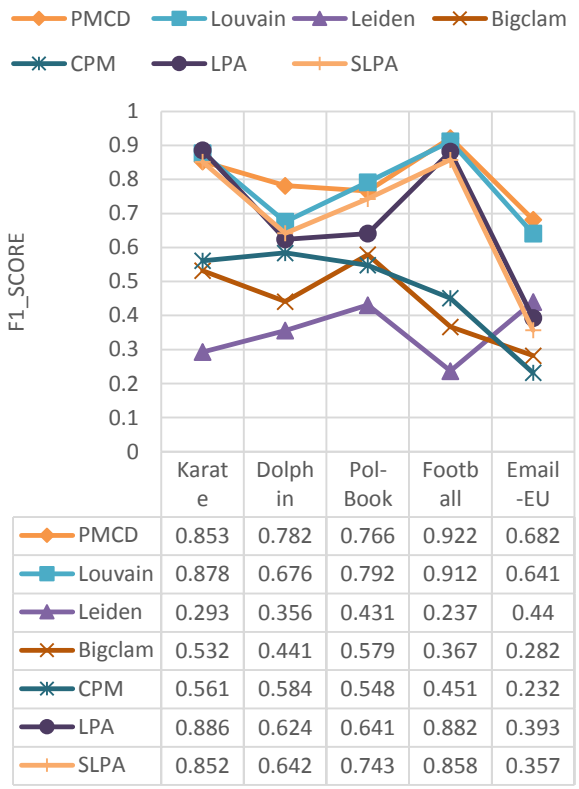


Fig. 4: F1_Score assessment chart, compare PMCD by community detection models on five real-world data sets.

Artificial Datasets

Utilizing an artificial network for evaluating community detection methods makes sense. Different methods can be used to generate artificial networks.

One of the most famous and often used strategies is the LFR benchmark [57]. The LFR benchmark builds types of artificial graphs with ground truth communities using density and dimension of communities. The network and community parameters can be set up before using LFR to simulate networks. The mixing parameter (μ) is one of the essential LFR parameters. This variable controls how various communities interact. A high mixing parameter value (μ), as indicated in Table 5, will reduce the network's level of modularity (Q_{GT}). As a result, the LFR-generated datasets are divided into two categories based on the modularity measure and mixing parameter: sparse communities and dense communities. The average degree is another crucial element that might be raised to encourage more intercommunal interaction.

Table 4 displays the key characteristics of the LFR artificial networks. Table 5 details the dataset that was created using our LFR approach.

Table 4: Parameters of LFR synthetic datasets [57]

Parameter	Description
N	Node number
K	Average degree
Min K	Minimum nodes degree
Max K	Maximum nodes degree
μ	Mixing parameter for the structure
Min C	Minimum for the community sizes
Max C	Maximum for the community sizes
tau1(γ)	The degree distribution
tau2(β)	The community size distribution

Table 5: The LFR artificial network properties

Graph Name	*N*	*k*	γ	β	μ	Q_{GT}
LFR-1	1000	20	3	1.5	0.05	0.895
LFR-2	1000	20	3	1.5	0.10	0.844
LFR-3	1000	20	3	1.5	0.15	0.800
LFR-4	1000	20	3	1.5	0.20	0.739
LFR-5	1000	20	3	1.5	0.25	0.699
LFR-6	1000	20	3	1.5	0.30	0.647
LFR-7	1000	20	3	1.5	0.35	0.603
LFR-8	1000	20	3	1.5	0.40	0.560
LFR-9	1000	20	3	1.5	0.45	0.504
LFR-10	1000	20	3	1.5	0.50	0.460
LFR-11	1000	20	3	1.5	0.55	0.407
LFR-12	1000	20	3	1.5	0.60	0.364
LFR-13	1000	20	3	1.5	0.65	0.321
LFR-14	1000	20	3	1.5	0.70	0.275
LFR-15	1000	20	3	1.5	0.75	0.229
LFR-16	1000	20	3	1.5	0.80	0.182

Table 6: Experimental results on sixteen LFR artificial networks by the modularity metric

Mixing Parameter (μ)	Louvain	Leiden	Bigclam	CPM	LPA	SLPA	PMCD
0.05	1.00	0.64	0.89	0.84	0.99	1.00	1.00
0.10	0.99	0.51	0.86	0.81	0.97	0.98	0.97
0.15	0.96	0.42	0.77	0.72	0.93	0.95	0.98
0.20	0.93	0.41	0.73	0.66	0.88	0.87	0.91
0.25	0.89	0.37	0.69	0.57	0.83	0.85	0.91
0.30	0.86	0.23	0.59	0.53	0.76	0.79	0.84
0.35	0.82	0.22	0.57	0.43	0.69	0.72	0.81
0.40	0.79	0.19	0.47	0.31	0.52	0.64	0.72
0.45	0.70	0.14	0.31	0.29	0.43	0.56	0.73
0.50	0.53	0.12	0.25	0.24	0.41	0.48	0.66
0.55	0.50	0.09	0.19	0.22	0.36	0.33	0.52
0.60	0.44	0.05	0.12	0.14	0.28	0.29	0.39
0.65	0.37	0.04	0.08	0.07	0.24	0.18	0.38
0.70	0.29	0.01	0.05	0.03	0.15	0.14	0.30
0.75	0.21	0.00	0.03	0.01	0.11	0.09	0.24
0.80	0.15	0.00	0.01	0.00	0.08	0.05	0.18

Experimental on Artificial Networks

In addition to actual graphs, we have also examined LFR artificial networks. We contrast the PMCD by the famous community detection model in Table 1 to demonstrate the result of the optimized loss function for a probabilistic estimate on the community detection utilizing modularity, F1_Score, and NMI measure. For this, sixteen LFR artificial networks are developed with various configures of mixing parameters (μ) ranging from 0.05 to 0.8, as indicated in Table 5. These networks are created in accordance with the attributes of synthetic networks listed in Table 4. Table 6's experimental findings demonstrate that the communities are dense for low mixing parameter range (e.g., 0.05≤μ≤0.4) and that the compared methods are almost correct in this situation.

However, the major contrast between the methods becomes more apparent when the mixing parameter's (μ) value rises (e.g., 0.4<μ≤0.8) and the communities are sparse, making it difficult to identify communities since the edges between communities' rise.

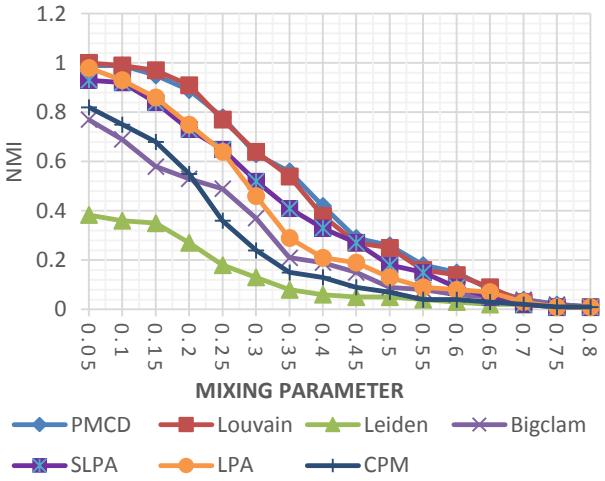


Fig. 5: NMI assessment graph on sixteen LFR datasets, comparing PMCD with six community detection methods.

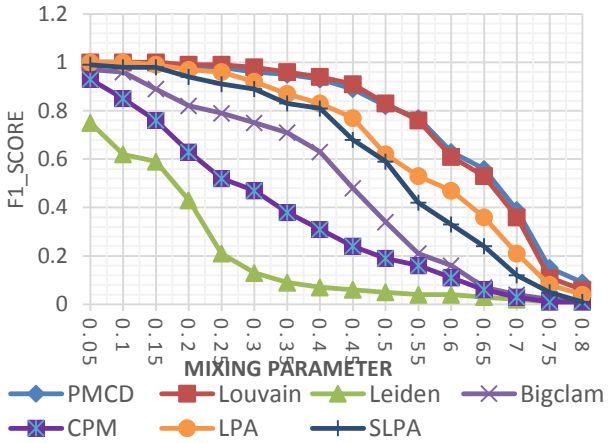


Fig. 6: On sixteen LFR datasets, the F1_score assessment chart compares PMCD with six community detection methods.

As can be seen in Figs 5 and 6, when the mixing parameter value increases, certain algorithms have NMI and F1Score values that are equivalent to zero. The majority of frequently used approaches in the range of 0.5 to 0.8 are outperformed by the suggested method.

Conclusion

We proposed a probabilistic motif-based method for detecting communities in complex networks. Due to the complexity of combining probabilistic approaches in motif structure, recent community detection methods have given the latent variable of the probabilistic model less consideration. However, the proposed approach leverages the intensity of the node's participation in the community and the relationship of at least two linked edges between three nodes (triangular motif structure) to estimate the hidden variable of the probabilistic model. The research maximized the likelihood function and extracted the model's latent parameters using the well-known block coordinate ascent technique. The association between node membership in communities and edge density is another aspect that helps in the examination of newly detected communities; three nodes are more likely to create a motif structure when seen in various communities. Overlapping in the identification of communities is another benefit of PMCD; according to the findings, communities that overlap have a greater density of triangular motifs. We employed 16 artificial graphs and 5 real graphs to evaluate the performance of the suggested method. In comparison to the other six methods, PMCD was able to achieve a sufficient quorum on real-world networks and surpass them in terms of internal and external assessment criteria. Synthetic network assessments further show that the suggested strategy performs better in sparse datasets than other approaches. Furthermore, a review of the complexity of the execution time reveals that the suggested method outperforms previous approaches. Future research can develop PMCD. A probabilistic generative model can be used to estimate edge weight while taking a latent parameter into account. Also, the suggested method can be enhanced by utilizing network node properties in order to give a more precise description of the found communities.

Author Contributions

H. Hajibabaei, V. Seydi, and A. Koochari contributed to the research design and implementation, the analysis of the results, and the writing of the manuscript.

Acknowledgment

We thank the editor and all the anonymous reviewers.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy, have been completely observed by the authors.

Abbreviations

PMCD	probabilistic motif-based community detection
NMI	Normalized Mutual Information
LFR	Lancichinetti–Fortunato–Radicchi Benchmark
CPM	Clique Percolation Method
LPA	Label Propagation Algorithm
SLPA	Speaker-Listener Label Propagation Algorithm
Q_{GT}	Ground Truth Modularity
SNAP	Stanford Network Analysis Project

References

[1] J. Sia, E. Jonckheere, P. Bogdan, "Ollivier-ricci curvature-based method to community detection in complex networks," Sci. Rep., 9(1): 1-12, 2019.

[2] Y. Y. Ahn, J. P. Bagrow, S. Lehmann, "Link communities reveal multiscale complexity in networks," Nature, 466(7307): 761-764, 2010.

[3] J. J. McAuley, J. Leskovec, "Learning to discover social circles in ego networks," in Proc. NIPS: 548-556, 2012.

[4] M. Girvan, M. E. Newman, "Community structure in social and biological networks," PNAS, 99(12): 7821-7826, 2002.

[5] W. Wu, S. Kwong, Y. Zhou, Y. Jia, W. Gao, "Nonnegative matrix factorization with mixed hypergraph regularization for community detection," Inf. Sci., 435: 263-281, 2018.

[6] S. Fortunato, D. Hric, "Community detection in networks: A user guide," Phys. Rep., 659: 1-44, 2016.

[7] H. Hajibabaei, V. Seydi, A. Koochari, "Community detection in weighted networks using probabilistic generative model," J. Intell. Inf. Syst., 60: 119- 136, 2023.

[8] T. S. Wang, H. T. Lin, P. Wang, "Weighted-spectral clustering algorithm for detecting community structures in complex networks," Artif. Intell. Rev., 47(4): 463-483, 2017.

[9] X. Chen, J. Li, "Community detection in complex networks using edge-deleting with restrictions," Physica A, 519: 181-194, 2019.

[10] F. D. Zarandi, M. K. Rafsanjani, "Community detection in complex networks using structural similarity," Physica A, 503: 882-891, 2018.

[11] B. D. Le, H. Shen, H. Nguyen, N. Falkner, "Improved network community detection using meta-heuristic based label propagation," Appl. Intell., 49(4): 1451-1466, 2019.

[12] C. Lyu, Y. Shi, L. Sun, "A novel local community detection method using evolutionary computation," IEEE Trans. Cybern., 51(6): 3348-3360, 2019.

[13] W. Zhou, X. Wang, C. Zhang, R. Li, C. Wang, "Community detection by enhancing community structure in bipartite networks," Mod. Phys. Lett. B, 33(7): 1950076, 2019.

[14] J. Yang, J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in Proc. the Sixth ACM International Conference on Web Search and Data Mining: 587-596, 2013.

[15] T. Ma *et al.*, "LED: A fast overlapping communities detection algorithm based on structural clustering," Neurocomputing, 207: 488-500, 2016.

[16] F. Liu, D. Choi, L. Xie, K. Roeder, "Global spectral clustering in dynamic networks," PNAS, 115(5): 927-932, 2018.

[17] J. Yang, J. Leskovec, "Community-affiliation graph model for overlapping network community detection," in Proc. 2012 IEEE 12th International Conference on Data Mining: 1170-1175, 2012.

[18] G. Palla, I. Derényi, I. Farkas, T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," nature, 435(7043): 814-818, 2005.

[19] X. Zhou, K. Yang, Y. Xie, C. Yang, T. Huang, "A novel modularity-based discrete state transition algorithm for community detection in networks," Neurocomputing, 334: 89-99, 2019.

[20] M. E. Newman, M. Girvan, "Finding and evaluating community structure in networks," Phys. Rev. E, 69(2): 026113, 2004.

[21] K. Berahmand, A. Bouyer, "A link-based similarity for improving community detection based on label propagation algorithm," J. Syst. Sci. Complexity, 32(3): 737-758, 2019.

[22] U. N. Raghavan, R. Albert, S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," Phys. Rev. E, 76(3): 036106, 2007.

[23] M. Zarezade, E. Nourani, A. Bouyer, "Community detection using a new node scoring and synchronous label updating of boundary nodes in social networks," J. AI Data Min., 8(2): 201-212, 2020.

[24] S. Kumar, B. Panda, D. Aggarwal, "Community detection in complex networks using network embedding and gravitational search algorithm," J. Intell. Inf. Syst., 57: 51-72, 2021.

[25] A. Torkaman, K. Badie, A. Salajegheh, M. H. Bokaei, S. F. Fatemi, "A hybrid deep network representation model for detecting researchers' communities," J. AI Data Min., 10(2): 233-243, 2022.

[26] X. Su *et al.*, "A comprehensive survey on community detection with deep learning," IEEE Trans. Neural Networks Learn. Syst., 2022.

[27] M. Ali, M. Hassan, K. Kifayat, J. Y. Kim, S. Hakak, M. K. Khan, "Social media content classification and community detection using deep learning and graph analytics," Technol. Forecasting Social Change, 188: 122252, 2023.

[28] C. Li, Y. Tang, Z. Tang, J. Cao, Y. Zhang, "Motif-based embedding label propagation algorithm for community detection," Int. J. Intell. Syst., 37(3): 1880-1902, 2022.

[29] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, A. Baig, "Community detection in networks: A multidisciplinary review," J. Network Comput. Appl., 108: 87-111, 2018.

[30] K. Guo, X. Huang, L. Wu, Y. Chen, "Local community detection algorithm based on local modularity density," Appl. Intell., 52(2): 1238-1253, 2022.

[31] J. Sánchez-Oro, A. Duarte, "Iterated Greedy algorithm for performing community detection in social networks," Future Gener. Comput. Syst., 88: 785-791, 2018.

[32] V. A. Traag, L. Waltman, N. J. Van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," Sci. Rep., 9(1): 1-12, 2019.

[33] C. Li, H. Chen, T. Li, X. Yang, "A stable community detection approach for complex network based on density peak clustering and label propagation," Appl. Intell., 52(2): 1188-1208, 2022.

[34] J. Xie, B. K. Szymanski, X. Liu, "Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in Proc. 2011 ieee 11th International Conference on Data Mining Workshops: 344-349, 2011.

[35] S. Gregory, "Finding overlapping communities in networks by label propagation," New J. Phys., 12(10): 103018, 2010.

[36] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, T. Vicsek, "CFinder: locating cliques and overlapping modules in biological networks," Bioinformatics, 22(8): 1021-1023, 2006.

[37] P. Bloem, S. de Rooij, "Large-scale network motif analysis using compression," Data Min. Knowl. Discovery, 34: 1421-1453, 2020.

[38] A. Arenas, A. Fernandez, S. Fortunato, S. Gomez, "Motif-based communities in complex networks," J. Phys. A: Math. Theor., 41(22): 224001, 2008.

[39] C. E. Tsourakakis, J. Pachocki, M. Mitzenmacher, "Scalable motif-aware graph clustering," in Proc. the 26th International Conference on World Wide Web: 1451-1460, 2017.

[40] L. Huang, H. Y. Chao, Q. Xie, "MuMod: A micro-unit connection approach for hybrid-order community detection," in Proc. the AAAI conference on artificial intelligence, 34(01): 107-114, 2020.

[41] J. Yang, J. McAuley, J. Leskovec, "Community detection in networks with node attributes," in Proc. IEEE 13th International Conference on Data Mining: 1151-1156, 2013.

[42] K. Yang, Q. Guo, J. G. Liu, "Community detection via measuring the strength between nodes for dynamic networks," Physica A, 509: 256-264, 2018.

[43] W. Yu, W. Wang, P. Jiao, X. Li, "Evolutionary clustering via graph regularized nonnegative matrix factorization for exploring temporal networks," Knowledge-Based Syst., 167: 1-10, 2019.

[44] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, "Network motifs: simple building blocks of complex networks," Science, 298(5594): 824-827, 2002.

[45] Y. Xu, W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," SIAM J. Imag. Sci., 6(3): 1758-1789, 2013.

[46] C. J. Hsieh, I. S. Dhillon, "Fast coordinate descent methods with variable selection for non-negative matrix factorization," in Proc. the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 1064-1072, 2011.

[47] C. J. Lin, "Projected gradient methods for nonnegative matrix factorization," Neural Comput., 19(10): 2756-2779, 2007.

[48] D. F. Gleich, C. Seshadhri, "Vertex neighborhoods, low conductance cuts, and good seeds for local community methods," in Proc, the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 597-605, 2012.

[49] V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, "Fast unfolding of communities in large networks," J. Stat. Mech: Theory Exp., 2008(10): P10008, 2008.

[50] A. Clauset, M. E. Newman, C. Moore, "Finding community structure in very large networks," Phys. Review E, 70(6): 066111, 2004.

[51] W. W. Zachary, "An information flow model for conflict and fission in small groups," J. Anthropol. Res., 33(4): 452-473, 1977.

[52] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," Behav. Ecol. Sociobiol., 54(4): 396-405, 2003.

[53] J. Kunegis, "Konect: the koblenz network collection," in Proc. the 22nd International Conference on World Wide Web: 1343-1350, 2013.

[54] J. Leskovec, J. Kleinberg, C. Faloutsos, "Graph evolution: Densification and shrinking diameters," ACM Trans. Knowl. Discovery Data (TKDD), 1(1): 2-es, 2007.

[55] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 555-564.

[56] R. Rossi, N. Ahmed, "The network data repository with interactive graph analytics and visualization," in Proc. Twenty-ninth AAAI Conference on Artificial Intelligence, 2015.

[57] A. Lancichinetti, S. Fortunato, F. Radicchi, "Benchmark graphs for testing community detection algorithms," Phys. Review E, 78(4): 046110, 2008.

Biographies

Hossein Hajibabaei received his M.Sc. in Software Engineering in 2014. Currently, he is a Ph.D. candidate in Computer Engineering majoring in artificial intelligence, and is teaching as a teaching assistant at the Islamic Azad University of Science and Research Branch. His areas of interest are social network analysis and deep learning.

- Email: h.hajibabaei@srbiau.ac.ir
- ORCID: 0009-0005-8063-981X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Vahid Seydi is a Senior Research Fellow in the School of Ocean Science at Bangor University in Data Science (DS) and Machine Learning (ML). Before Bangor, He was an Assistant Professor at the Department of Al at Azad University South Tehran Branch (Feb 2014 - Sep 2020) and an award-winning lecturer (Oct 2010 Feb 2014). Vahid received a B.Sc. (2005) in software engineering, and M.Sc. (2007), and Ph.D. (2014) in Al, from the Department of Computer Science at Azad University, Science and Research Branch, Tehran Iran. He has been awarded a Global Talen endorsement from the UK Royal Society (2023), His current research fellowship (2020), and a merit-based scholarship for attending the school of Al, Rome, Italy (2019), Also, He has achieved a full scholarship Award from Azad University (2010-2014), KNTU ISLAB Research Fellowship (2007-2010). He secured the first rank among the graduates from 2004-2005. His current research focuses on dedicating machine learning methods to analyze data associated with digital oceanography, especially in the offshore renewable energy section.

- Email: vahidseydi@gmail.com
- ORCID: 0000-0001-5702-2209
- Web of Science Researcher ID: NA
- Scopus Author ID: 23490316700
- Homepage: https://vahidseydi.github.io

Abbas Koochari received his Ph.D. in Computer Engineering majoring in artificial intelligence. He is currently an assistant professor and a member of the scientific staff of Islamic Azad University, Science and Research Branch. His areas of interest are image processing, machine vision, speech and natural language processing, and deep learning.

- Email: koochari@srbiau.ac.ir
- ORCID: 0000-0003-0584-6470
- Web of Science Researcher ID: NA
- Scopus Author ID: 36005396600
- Homepage: https://srb.iau.ir/faculty/a-koochari/en

How to cite this paper:
H. Hajibabaei, V. Seydi, A. Koochari, "Motif-based community detection: A probabilistic model based on repeating patterns," J. Electr. Comput. Eng. Innovations, 12(1): 247-258, 2024.

DOI: 10.22061/jecei.2023.9931.663

URL: https://jecei.sru.ac.ir/article_2013.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Applying Partial Differential Equations on Cubic Uniform Local Binary Pattern to Reveal Micro-Changes

V. Esmaeili [1], M. Mohassel Feghhi [1,], S. O. Shahdi [2]*

[1] *Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran.*

[2] *Department of Electrical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran.*

Article Info

Article History:
Received 23 August 2023
Reviewed 12 October 2023
Revised 07 November 2023
Accepted 04 December 2023

Keywords:
Apex
Cubic-ULBP
Micro-Changes
Micro-Expression
Partial Differential Equations (PDE)

*Corresponding Author's Email Address:
mohasselfeghhi@tabrizu.ac.ir

Abstract

Background and Objectives: The world we live in everyday, accompany with enormous numbers of minute variations which affect us and our surroundings in several aspects. These variations, so called micro-changes, are low in intensity and brief in duration which makes them almost invisible to naked eyes. Nonetheless, revealing them could open up a new wide range of applications from security, business, engineering, medical, and seismology to psychology.

Methods: In this paper, we adopted a novel autonomous approach comprising Partial Differential Equations (PDE) and Cubic Uniform Local Binary Pattern (Cubic-ULBP) to spot micro-changes. Cubic-ULBP extracts 15 planes containing robust features against noise and illumination variations. Afterwards, PDE pick out single plane out of 15 to reduce time consumption.

Results: The proposed method is not only optimized to get the job done perfectly but also provides promising results comparing with most state-of-the-art methods. So that the accuracy is increased about 36% and 40% on the CASME and the CASME II databases, respectively.

Conclusion: The combination of the PDE and the Cubic-ULBP creates a strong and optimal method for detecting the apex frame and micro-movement. This method's performance is found to be promising when the training samples are scarce, too.

Introduction

The micro-changes take place frequently in various environments, which their existence could be confirmed through elaborate analysis of the video frames. Meanwhile, the human eyes cannot see such trifle changes since they appear in fraction of second/s without clear perceptible features. Undoubtedly, spotting them is advantageous to many applications such as security, engineering, medical, judicial, nanotechnology seismology, psychology, and etc.

Ekman and Friesen are believed to be the first researchers who unveil facial micro-changes when they were scrutinizing video frames in 1969 [1]. These micro-changes appear on face both instantaneous and spontaneous against person willingness to suppress them. Besides, their characteristics make micro-changes, known as micro expression, genuine and real [2]-[3]. Hence, such nonverbal leakage could be regarded as a strong clue for the act of lying and deception [1], [4]-[8].

Although spotting such a low intensity and moment facial muscle deformity is not an easy task, revealing them paves the long way for court authorities to discern honest

from lying culprits and/or witnesses. Moreover, with these micro-expression symptoms, psychologists can early diagnose patients with depression and other mental health difficulties. As a result, it could play a vital role to avoid suicide or related mental disease aftermath. In the same manner, within educational systems, teachers could measure their training efficacy with extracting student micro-expressions.

The micro-changes are not limited to micro expression, but to other key areas. In another case, we could locate micro-change features when smoke is arisen above the mouth of a volcano. The amount of smoke can determine the rate of volcano activity [9], [10]. Surprisingly, one could even foresee the earthquake by timely detecting subtle and micro smoke changes [11], [12]. Therefore, casualties caused by great destruction due to volcanic actions are diminished significantly, and human lives are saved.

In more direct connection with human health, spotting micro-changes in blood circulation can aid diagnosing some of the most prevalent medical diseases. As an example, a cardiac arrhythmia is high probable when person has an irregular heartbeat, i.e. too slow, or too fast [13], which could be stemmed from heart disease [14]. Also, monitoring heartbeat or pulse can determine palpitations and locate local blood vessels. In fact, their visibility comes from micro-changes of blood which flows in arteries closed to the surface of skin.

In retrospect, detection of such micro-changes could result in making concise decision for wide range of applications. Nevertheless, this task is an arduous and requires professional expertise or autonomous setup. One could spot micro-changes with high attention over time well. For further clarification, assume a phenomenon including micro-changes. When each image has taken into probe individually, they look all the same. However, when we compare them together over time, their minute disparities would be discovered.

This minute scale changes, which are occurred in various locations and multi-orientations within image frames, impose severe detection challenge. Their subtle changes are indiscernible from shadows, noises, and other similar features. As a result, the performance of spotting micro-changes is subsided significantly with the presence of noise or equally by low-rate of the signal to noise ratio (S/N).

It is noteworthy to mention that among the sequence of frames; one single frame has the highest fluctuations in features with respect to others. This frame is known as the peak or apex, which is very crucial to locate. It can convey the strongest signal and main message. For instance, a single apex represents vital information with the video capture of micro-expressions at 200 fps. It contains sufficient meaningful information which could be then interpreted as a specific micro-expression [15], [16]. The rest of the frames demonstrate nearly a neutral expression of the face, or less deviated from neutral compare to apex. Hence, analyzing all of the frames rather than just apex increases the burden of processing without a fruitful outcome.

The apex frame, also in the case of volcano, contains the maximum smoke coming out of a crater which could be used to approximate the time of probable earthquakes. In another case, the apex frame makes it easy to locate blood vessels under the skin. This process alleviates injecting donated blood, blood products, or other necessary fluids into the circulatory system.

Since the direction of micro-change is unpredictable, all possible orientations must be considered. Thus, promising textural algorithms, like Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) [17], are not satisfying for the micro-changes analysis. They are imperfect since they merely take data in two directions (i.e. vertical and horizontal) over time.

Instead of three planes in the limited directions, the Cubic-LBP [4] comprising 15 planes in multiple directions which potentially grasp any changes in multi-directions. Nevertheless, both the LBP-TOP and the Cubic-LBP adversely affected by illumination variations and noises as they exert the simple LBP [18] on their planes. Using Uniform LBP (ULBP) [19] rather than the simple one can discard the interference of illumination variations and shadows in micro-change spotting process. Also, processing 15 planes in Cubic-LBP has more computation. Therefore, it is very time-consuming.

In 2020, Esmaeili and shahdi [4] tried to find the apex frame containing the micro-changes in the CASME II and the CASME. In their work, the error was about 6.5 frames from the truth apex in the CASME II. Furthermore, there was around two frames error in the CASME. Then, in [6], the use of the diagonal planes not only reduced the error but also led to simple calculations. However, there were still six frames far away from the ground truth in the CASME II. In the next work [20], a method called the LBP on six intersection planes had only 1.7 errors in the CASME.

In 2022, a method based on the deep learning named as the intelligent cubic-LBP [21] has been presented for spotting the micro-changes on the datasets such as the FA and the POVI. Nevertheless, the convolutional neural networks often require complex computation and the high processing/learning time [22]-[26].

In this paper, we present a novel approach to detect the micro-changes. Our goal is to bring the Cubic-ULBP and the Partial Differential Equations (PDE) together for the sake of reaching outstanding outcomes. In fact, the major contribution of this paper is to extract robust features against noises through the Cubic-ULBP.

Furthermore, the proposed method automatically selects only one plane out of the 15 Cubic-LBP planes using the PDE, to reduce the processing time. This single plane highlights the main direction of the micro-changes. Our extensive experiments on datasets prove the competitiveness of the proposed method, which is optimized to contests with the related state-of-the-arts.

Related Works

A plethora of phenomena is tending to cause changes, movements, and motions that are still subtle in nature [27]. Even though, they are too small to distinguish by the human visual system accurately. In order to address this limitation, researchers have suggested various methods to tackle this problem which ends up with detecting minute spatial and temporal variations [4], [28]-[30].

The Lagrangian method is considered to be one of them. It tracks movement or motion in a video frames [28], [31]. The principle of this method is based on magnifying movements of fine point which consequently become more visible to the naked eye. However, poor estimated movements and sensitivity to spatial noise are among its drawbacks.

The other motion magnification method is called Eulerian that is used in multiple fields [29], [32]-[34]. Beside magnification, it also reveals micro-color changes [28]. Deteriorating by captured noise, computationally-intensive, and unsuitability for high-frequency bands are its general disadvantages. Unlike Lagrangian, the Eulerian doesn't support large amplification factors.

On the contrary, phase-based video motion processing approach [27] is robust against noise due to phase magnification rather than feature magnification. However, because of its complexity, the huge processing burden is imposed on such system.

The other intricate method is the complex wavelet (CWT), which is optimized to magnify micro-movement [29]. Under this approach, any micro-change in any direction over time is magnified by the CWT decomposition band.

Two later methods mentioned above rely on the frequency decomposition.

To resolve the complexity problem, LBP-based methods such as Cubic-LBP [4] have been introduced. Nonetheless, they have deficiencies that should be get rid of them.

These are sensitive to the illumination variations, the shadow variations, and the noises. It is in order to come up with aforementioned shortcomings, we adopt an optimized approach comprising Cubic-ULBP and PDE.

The Cubic-ULBP is not only simple in computation but also can track any given direction of the micro-change. On top of that, it is insensitive to noise and illumination variations.

Unfortunately, the Cubic-ULBP has relatively high number of planes, i.e. 15 planes. To eliminate this problem, the PDE is then called to choose just single plane. Therefore, volume of computation becomes even simpler. It is worthwhile to bear in mind that PDE has other applications in image processing such as denoising, image enhancement, and tracking as reported in [35]-[40].

Proposed Method

We extract features from the consecutive images using a texture-based method, namely Cubic-ULBP. Then, the maximum direction of micro-changes would be determined.

More specifically, the PDE selects a single plane among 15 that reveals the most micro-changes. Finally, the apex frame is spotted by evaluating the discrepancy between the histogram of each frame and the normal frame in that single plane.

The general framework of our proposed method is shown in Fig. 1.

A. Cubic-ULBP

We use frames as image sequences. These standard images are put successively such that construct a three-dimensional array. 15 planes encompass the pixels in multi-directions. They are named Cubic method planes as depicted in Fig. 2.

There are 3 separate planes in directions of XT, XY, and YT.

Also, there are 6 diagonal planes in the cuboid. Except for the XY orientation planes, the rest of them collect pixels from consecutive images. It is prerequisite to convert the color images to the grayscale for ULBP process.

Suppose pixels ($P_{\S}$, for *§=1* to *NuP*) are located on a circle with Ra radius in a plane. The number of $P_{\S}$ *(NuP)* is selected to attain the best performance.

If we consider that the middle point of the circle to be a pixel *C*, then, the location of each $P_{\S}$ on every plane is defined as follow:

$$\left(x_C + Ra_x\, cos(I)\,, y_C - Ra_y\, sin(I)\,, t_C\right) \quad (1)$$

$$\left(x_C + Ra_x cos(I), y_C - Ra_y sin(I), t_C \pm S\right) \quad (2)$$

$$\left(x_C, y_C - Ra_y sin(I), t_C + Ra_t cos(I)\right) \quad (3)$$

$$\left(x_C \pm Ra_x, y_C - Ra_y sin(I), t_C + Ra_t cos(I)\right) \quad (4)$$

$$\left(x_C + Ra_x cos(I), y_C, t_C + Ra_t sin(I)\right) \quad (5)$$

$$\left(x_C + Ra_x cos(I), y_C \pm Ra_y, t_C + Ra_t sin(I)\right) \quad (6)$$

$$\left(x_C + Ra_x sin(I), y_C - Ra_y sin(I), t_C - Ra_t cos(I)\right) \quad (7)$$

$$\left(x_C - Ra_x sin(I), y_C - Ra_y sin(I), t_C + Ra_t cos(I)\right) \quad (8)$$

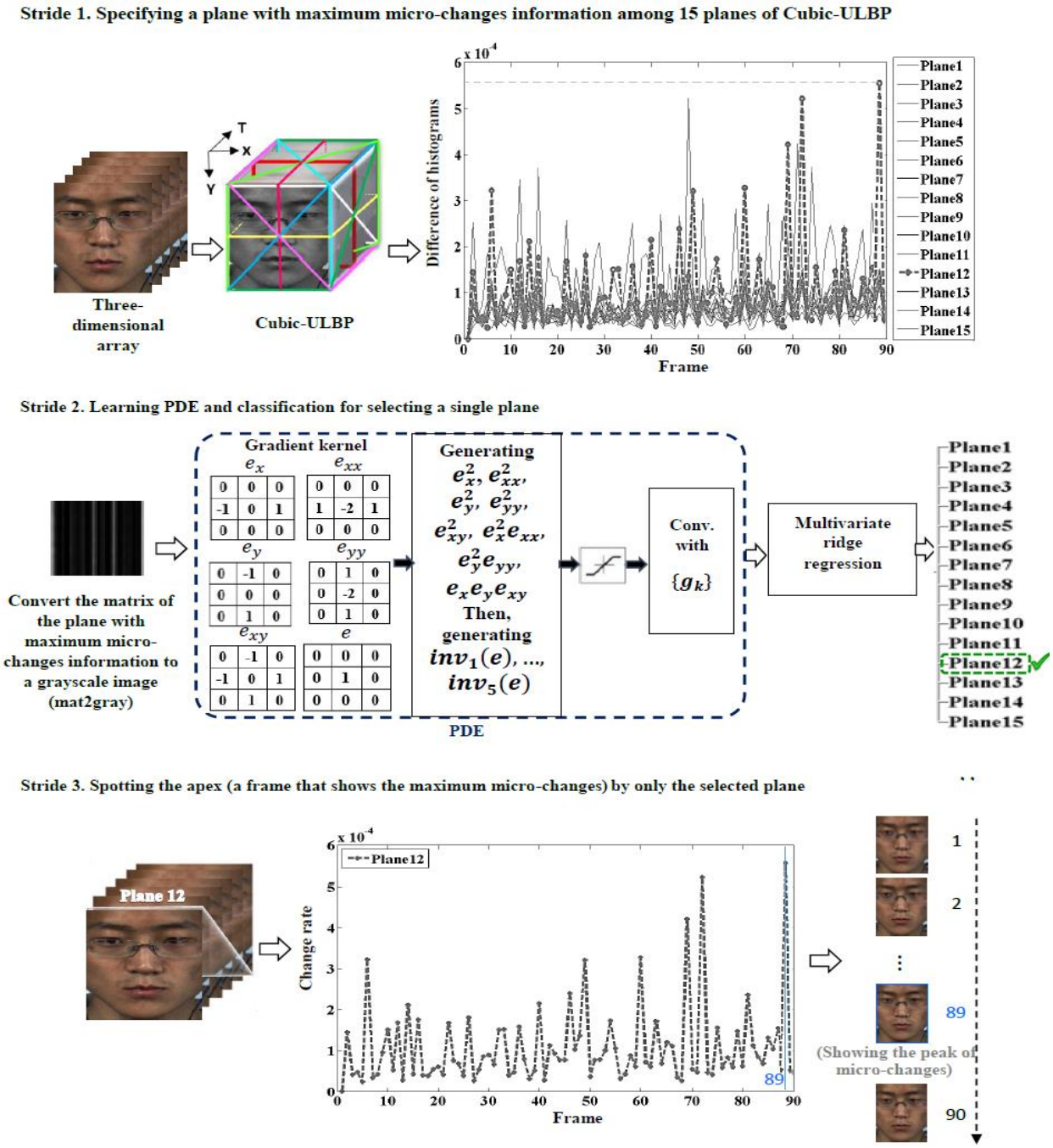


Fig. 1: The general framework of our proposed method.

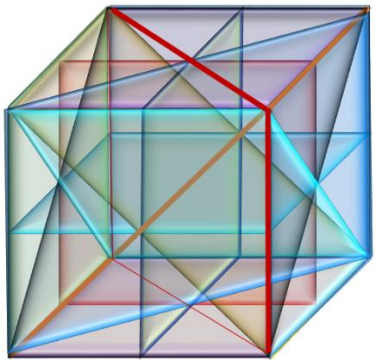


Fig. 2: Cubic method planes.

$$\left(x_C + Ra_x cos(I), y_C - Ra_y sin(I), t_C + Ra_t sin(I)\right) \quad (9)$$

$$\left(x_C - Ra_x cos(I), y_C - Ra_y sin(I), t_C - Ra_t sin(I)\right) \quad (10)$$

$$\left(x_C + Ra_x cos(I), y_C - Ra_y sin(I), t_C - Ra_t cos(I)\right) \quad (11)$$

$$\left(x_C + Ra_x cos(I), y_C - Ra_y sin(I), t_C + Ra_t cos(I)\right) \quad (12)$$

where *I* equal to *2πP/NuP* and *S* is the interval on *T* axis.

In the ULBP routine, $P_{\acute{s}}$ values are subtracted by the value of *C*. Then, the positivity of outcome is checked. If it holds true, it would be exchanged by "1" and otherwise

replaced by "0". Thereafter, these binary values concatenated together to create specific pattern (e.g. 01101001).

In order to discern uniform from non-uniform patterns, we build a histogram of ULBP. For this purpose, number of reciprocal transitions between 0 and 1 are enumerated. If there are just 2 or less transitions, the pattern is put into uniform bin. A separate bin is assigned for all non-uniform patterns which have more than two transitions. This task provides the feature vector, well known as histogram of ULBP.

In comparison with ordinary LBP [18], ULBP ends up in shorter feature vector. It could be obtained through following equation:

$$e_q(x,y,t) = 0, \quad (x,y,t) \in \partial\Psi \times [0,Ti]$$

$$Uniform\ LBP = |sign(Val_{NuP} - Val_C) - sign(Val_0 - Val_C)| + \sum_{P=1}^{NuP-1} |sign(Val_P - Val_C) - sign(Val_{P-1} - Val_C)| \quad (13)$$

$$HistULBP = \begin{cases} \sum_{P=0}^{NuP-1} sign(Val_P - Val_C) & if\ Uniform\ LBP_{NuP,Ra} \le 2 \\ NuP + 1 & otherwise \end{cases} \quad (14)$$

To get a coherent description, the normalized histogram of the Cubic-ULBP is computed by concatenation of ULBP features on the planes as follow:

$$NormHist_{\varphi,h} = \sum_{x,y.t} J\{Uniform\ LBP_{\varphi}(x,y,t) = \varphi\} / \sum_{\omega=0}^{\varrho-1} HistULBP_{\omega,h} \quad (15)$$

where φ is a factor which is confined between 0 and the number of ULBP labels in the plane (ϱ). The h is plane number (i.e. *h: 1, 2, …, 15*). *J (.)* is 1 when (.) is true, and it is 0 otherwise.

The Cubic-ULBP histogram is comprised of 15 histograms; each one belongs to specific counterpart plane. In most cases, only one plane out of 15 planes carries information with the most micro-changes. It can be explained mathematically as:

$$\sum_1^{NF} (NormHist_{fi} - NormHist_f)^2 \quad (16)$$

where *fi*, *NF*, and *f* are the frame of interest, number of frames, and frame1 respectively. With frame1, we mean a normal frame, containing no change, which is often happened to be the first frame [41], [42].

B. PDE

Histogram of a plane with the most micro-changes is captured as a matrix ($\boldsymbol{v}$). The histogram, which is obtained from Cubic-ULBP in previous section, is processed again to create vector *v*. Thereupon, the *v* is inserted into the PDE and its features are extracted as output (*o*). The PDE is defined as:

$$\frac{\partial e_q}{\partial t} = L(e,x,y,t), \quad (x,y,t) \in \Psi \times [0,Ti]$$

$$e_q\big|_{t=0}(x,y,t) = \boldsymbol{v_q}, \quad (x,y) \in \Psi \quad (17)$$

where *q=1, 2, …, Q*. The *Q* is the number of samples and Ψ is the rectangular region in R^2. The *e* and *Ti* are the evolution of the *v* and feature extraction process time by the PDE respectively. The feature map $e \mid t = T$ has a similar dimension as the *v*. If the gradient and hessian of the *v* are $\nabla v\ and\ \boldsymbol{H}_v$, then the PDE can be formulated as:

$$\frac{\partial e}{\partial t} = L(e,\nabla \boldsymbol{v},\boldsymbol{H}_v) \quad (18)$$

We deduce the L feature inspired by [40]. The PDE shares the same characteristics with Cubic-ULBP which both are illumination and rotation invariant. In fact, we use the rotational invariants up to the second order. The input and desired output are as follows, respectively:

$$inv_0(\mathrm{e}) = 1 \quad (19)$$

$$inv_1(\mathrm{e}) = e \quad (20)$$

The squared norm of the gradient is as:

$$inv_2(\mathrm{e}) = \|\nabla e\|^2 = e_x^2 + e_y^2 \quad (21)$$

Also, the Laplacian is as follows:

$$inv_3(\mathrm{e}) = tr(\boldsymbol{H_e}) = e_{xx} + e_{yy} \quad (22)$$

A visual front-end operation is as:

$$inv_4(\mathrm{e}) = (\nabla e)^{Ti} \boldsymbol{H}_e \nabla e = e_x^2 e_{xx} + 2e_x e_y e_{xy} + e_y^2 e_{yy} \quad (23)$$

In addition, a deviation from flatness is as following:

$$inv_5(\mathrm{e}) = tr(\boldsymbol{H}_e^2) = e_{xx}^2 + 2e_{xy}^2 + e_{yy}^2 \quad (24)$$

To get approximate invariant under gray-level scaling, the term $a(x) = \frac{x}{1+|x|}$ is added to each differential invariant. Thus, the differential invariants becomes as $\{a(inv_{0\ to\ 5}(e))\}$. Consequently, *L* can be written as a function of them:

$$L(e,t) = \sum_{k=0}^{5} g_k(t)\ a(inv_k(e(t))) \quad (25)$$

$g_k(t)$ is independent of the (x, y) in (25). It is to minimize the loss function.

C. Classification

Now, a simple linear classifier is required for the classification purpose which is fed by PDE outcomes. We use the multivariate ridge regression [39], [40] in order to reduce complexity.

In the training step, we minimize a loss function to get optimized values for the parameters (*W*) and *L*. Training samples are the input matrixes ($\boldsymbol{v_q}$) and their

corresponding tag vector (d_q) that belongs to class h. For each input, we compute a feature map using (17). If the regularization term set to be G, then the whole learning features can be extracted as follow:

$$\underset{\{g_{k(t)}\},\boldsymbol{W}}{min}\, E = \frac{1}{Q}\sum_{q=1}^{Q} LOSS\left(\boldsymbol{W}; e_{q|t=Ti}, d_q\right) + \lambda\, G(\boldsymbol{W}) \quad (26)$$

where lambda (λ) is a trade-off parameter. The extracted features are chosen in order to minimize the loss function. More clearly, the PDE learns the $\{g_{k(t)}\}$ and W parameters. The aim in (26) is to learn the multivariate ridge regression as:

$$\underset{\boldsymbol{W},L}{min} E = \frac{1}{Q}\|D - \boldsymbol{W}.\boldsymbol{V}|_{t=Ti}\|_L^2 + \lambda\|\boldsymbol{W}\|_L^2 \quad (27)$$

where$D = \left[d_1, d_2, \ldots, d_Q\right]$, $and\ \boldsymbol{V}|_{t=Ti} = \left[vect(e_1|_{t=Ti}), \ldots, vect\left(e_Q|_{t=Ti}\right)\right]$. For the multivariate ridge regression, the size of the W matrix is the number of categories multiplied by the number of the input image pixels. The class label for each test image can be achieved as follows:

$$lable = \arg\max\{\boldsymbol{W}.vect(e|_{t=Ti})\} \quad (28)$$

To solve (27), $\{(\boldsymbol{v_q}, d_q)\}_{q=1}^{Q}$, λ, and the step size during optimization (μ) are selected as inputs. To achieve best overall results, $\Delta t, N, \beta, \varepsilon, \gamma, and\ \gamma_{max}$ are initialized to 0.5, 5, 0.95, 10-6, 1, and 10, respectively. Besides, Λ is initialized with each entry uniformly sampled from [−1, 1]. To accomplish classification task, we perform five steps while $\gamma \leq \gamma_{max}$ and $\|E^{\gamma} - E^{\gamma-1}\| > \varepsilon$:

1) Set $e_q^0 = \boldsymbol{v_q}$ and calculate $e_q^{\zeta+1} = e_q^{\zeta} + \Delta t \sum_{k=0}^{5} g_k^{\zeta}\, .a\left(inv_k\left(e_q^{\zeta}\right)\right), \zeta = 0,1, \ldots, N-1$;
2) Solve $\boldsymbol{W} = \underset{W}{\arg min}\, \frac{1}{Q}\, \|D - \boldsymbol{W}.\boldsymbol{V}^N\|^2 + \lambda\|\boldsymbol{W}\|_L^2 = D.(V^N)^{Ti}.[\boldsymbol{V}^N.(V^N)^{Ti} + \lambda Q\eta]^{-1}$;
$\eta \epsilon R^{\text{the pixel number of an image} \times \text{the pixel number of an image}}$
3) Update Λ by one gradient descent as $\left(g_k^{\zeta}\right)^{\gamma+1} = \left(g_k^{\zeta}\right)^{\gamma} - \mu \frac{\partial E^{\gamma}}{\partial\left(g_k^{\zeta}\right)^{\gamma}}$;
4) Update $\mu = \beta\mu$;
5) Update $\gamma = \gamma + 1$.

Experimental Results

All experiments are implemented in MATLAB 2020b under Windows 7 on a PC equipped with a 3.5GHz CPU and 8GB RAM. We begin this section with describing the datasets which are used in our research. Thereafter, the settings and also implementation details are provided to accommodate the necessity of procedure replication. In last, we present an extensive investigation on results which then being compared to the most state-of-the-art methods.

A. Datasets Specifications

The Chinese Academy of Sciences Micro-Expressions (CASME) [41] and its extended version, namely CASME II [42] are among most widely used datasets. One of particular reason behind this popularity is the liable labels including apex frame. Specially, these pre-labelled apex frames lead to a concise evaluation of our proposed method. The earlier dataset has more than 190 samples each one contains of dozens consecutive facial images. The latter, which is bigger in size, has more than 240 samples.

As it is appeared in Fig. 3, the apex frame demonstrates relatively high degree of variations in comparison with the other frames. More specifically, in this figure, the emotion of disgust is most discernable within forehead area in the apex frame.

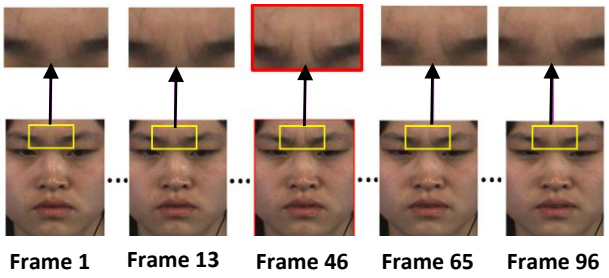


Fig. 3: A sample of disgust expression from the CASME II dataset. In this sample, the forty-sixth frame shows the most muscle contraction within the forehead area in comparison to others. Thus, it is the exact apex with the most micro-changes information.

Although the resolution of these datasets (1280×720 for CASME and 280×240 for CASME II) satisfies our system requirements, there are still some other motions which should not to be regarded as micro-changes. In order to overcome these defects, we manually omit the samples with the motions that are not mainly categorized as micro-changes. The third dataset which we manage to exploit in our proposed method is named POVI[1] [43]. It is video data which captured the activity of Villarrica volcano. A sample of image sequence from this dataset is illustrated in Fig. 4. It should be mentioned out that just the portion of this dataset which is in accordance with micro-changes of volcano activity is utilized.

We take advantage of the last dataset that is corresponding to retinal blood vessels. They were collected from the retinal Fluorescein Angiography (FA) [44]. For each instance in this dataset, there are at least 110 frames available. The main drawback with this dataset is the lack of apex frame labels which leaves researchers with no choice rather than tagging them manually.

[1] http://www.povi.cl/

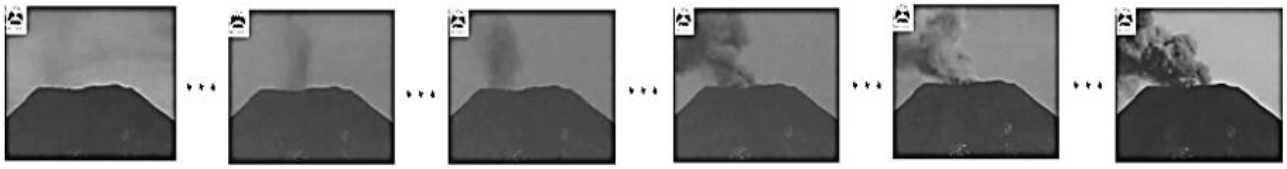


Fig. 4: A sample of consecutive images from the POVI.

B. Simulation Details

Initially, image frames of one sample are clustered with the length of three frames. This task proceeds by separating triple frames starting from onset frame into the offset one. To be sure not to miss any micro-change, this process is accomplished through an overlapping scheme where a frame could be considered multiple times. Thus, the interval on the T axis (*S*) is set to be 1. Besides, for the textural analyzer like ULBP to be applicable, all images should be first transformed into the grayscale level. As discussed in section III-A, cubes are made from clusters of 3 consecutive frames. Afterwards, 15 planes are extracted from each Cube to encompass every possible motion in multi-directions. It could be perceived from Fig. 5 that eight pixels including *P1, P2, …,* and *P8* are located on a circumference of a circle with a *Ra=1* radius. It means that the *NuP* is eight. The *Ra* and *NuP* have been selected according to the best performance. In order to apply ULBP, each pixel on the circle circumference is subtracted from the center *C* value. With this subtraction results, the string of binary codes is then created where positive and negative values of subtractions are exchanged with 1 and 0 respectively. If there are only 2 or less transition existed between zeros and ones within code string, this pattern is classified as uniform. Consequently, a separate bin is assigned for these uniform patterns in the ULBP histogram. The rest of the patterns, namely non-uniform patterns, are located into another single bin.

At last, all the histograms, which are extracted from 15 distinct planes, are concatenated to make the single Cubic-ULBP histogram. This histogram is then placed into a matrix of 15 rows in which each row stands for a specific plane. In other words, the ULBP histograms of plane 1 up to plane 15 are in turn the first row up to the last row of the matrix. Therefore, this matrix contains the entire textural features extracted from a sample video data by the means of Cubic-ULBP.

Now, we probe variations by looking into histogram matrix using (16). Thus, the plane with the maximum micro-changes is spotted when its variation is compared with the rest of the frames. This plane could pave the significant way to locate micro-changes since it has relatively high amount of valuable data. For example, plane number 13 in Fig. 5 has the same characteristic and outweigh in variations over other planes.

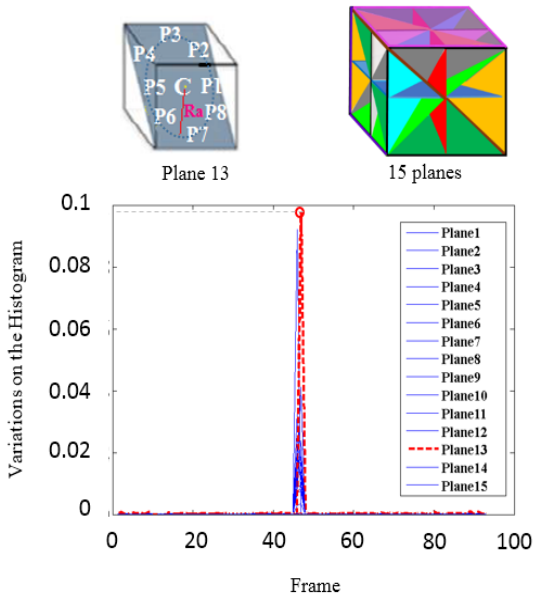


Fig. 5: The histograms variation using 15 planes in consecutive images.

In the next step, this plane's information is fed into the PDE, and in this example its corresponding tag is set to be 13. This procedure is iterated similarly for all other samples through the dataset.

The PDE parameters are trained to learn from features of the matrix according to part B of section III. We minimize the loss function for determining the parameters. The multivariate ridge regression is then utilized as a classifier. Training sets include histogram matrix as an input and its corresponding tag vector, which belongs to class h, as an output. As a result, there are 15 classes where the h1 as class 1 belongs to plane 1 and h15 as class 15 belongs to plane 15. Fortunately, the direction of micro-changes is also unraveled by this class.

For each plane, we randomly select 15 matrixes for training and the rest are selected for testing. In our method, the hyper-parameters of λ (i.e., regularization parameter) and μ (i.e., step size during optimization) are tuned to obtain the best performance. Since in most cases the training samples are limited, λ and μ are chosen from {1, 1.5, 5, 10, 50} and {0.1, 0.3, 0.5, 0.7, 0.9}, respectively [32]. We choose λ of 1.5 and μ of 0.5 for this purpose. We set them due to the best achieved spotting accuracy. Fig. 6 shows two samples of the effects of selecting λ and μ value on apex spotting accuracy on the used data.

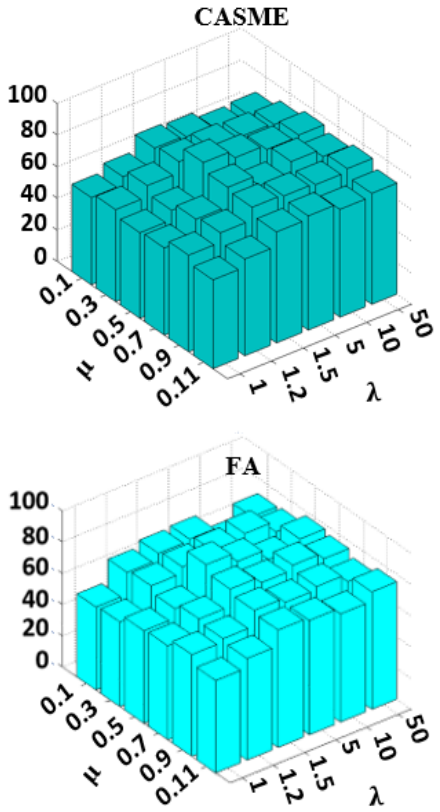


Fig. 6: The effects of λ and μ on apex spotting accuracy.

Therefore, the PDE pick out a single plane out of 15 that reveals the most micro-changes. Finally, the apex frame is spotted by evaluating the discrepancy between the histogram of each frame and the first frame (i.e., normal frame) in that single plane.

C. Discussion on Experimental Results

According to our results, the proposed method outperforms over other state-of-the-art researches in spotting the apex frame. Even more, its performance is found to be promising when the training samples are scarce. Meanwhile, the large amounts of training samples are often required for the well-known methods (such as the method presented in [21], [23], [26]). Collecting such enormous samples is not always easy and cost effective. Our proposed method not only overcomes this challenge, but also reaches better efficiency compared to the other approaches.

Besides, in terms of optimization, our proposed method has a clear advantage over customary Cubic-LBP since it selects merely a single plane rather than 15. In consequence, the computation process of our method is slashed significantly.

Unsurprisingly, with the characteristics of ULBP and PDE, our method is more robust against illumination variations. This robustness is mainly due to three reasons. First, we use a uniform pattern in ULBP that relatively discards effect of illumination variations. Second, the PDE parameters are learned with the features extracted from an illumination invariant histogram. Third, we add the term $a(x)$ on each differential invariant which ends up with more insensitivity to the illumination fluctuations.

Table 1 reports the apex spotting accuracy in percentage on the CASME dataset with 15 training samples. We can see that the apex frame spotting accuracy is increased by more than 35% in the experiments on the CASME in comparison with [2], [3], and [16], [17]. In order to further validate out method, the experiments are repeated on the CASME II.

Table 2 demonstrates micro-changes spotting accuracy on this dataset with just 15 training samples. Again, as it could be perceived from this table, the accuracy which is achieved by our method, in comparison with others, is improved noticeably.

In addition, we also conduct the experiments to come up with apex spotting in the POVI and FA. Their results are summarized in Table 3. Furthermore, the assessment of wrong apex spotting is computed through the mean absolute error as follow:

$$mean\ absolute\ error = \frac{1}{Q}\sum_{1}^{Q}|deviation| \tag{29}$$

Besides, we achieve the standard error as follow:

$$standard\ error\ = \frac{sample\ deviation}{\sqrt{Q}} \tag{30}$$

Results on both mean absolute error and standard error which are given in Table 4 are comparatively negligible. In Fig. 7 we picture a clear view of the superiority of our proposed method comparing with other related methods. Moreover, the proposed method (combining PDE and Cubic-ULBP) efficiency is measured using precision (prec.), recall (rec.), and F1 metrics [45] as follows:

$$rec.\ = \frac{TruePos.}{FalseNeg.+TruePos.} \tag{31}$$

$$prec.\ = \frac{TruePos.}{FalsePos.+TruePos.} \tag{32}$$

$$F1\ = 2 \times (\frac{prec.\times rec.}{prec.+rec.}) \tag{33}$$

where the Pos. and Neg. are abbreviations of positive and negative. Table 5 illustrates the above-mentioned metrics result. We have obtained high precision (0.86) using our proposed method.

In the last experiment that is reported in Table 6, we investigate the effect of different numbers of training samples on the accuracy. It is performed on CASME, CASME II, POVI, and FA which reaffirms the direct relationship between accuracy and numbers of training samples. However, our proposed method still keeps its performance even under severe reduction of training samples.

Finally, Fig. 8 compares the time consumption results with existing literature. According to the results, our

proposed method has less elapsed time than [4] and [21] works. Since the proposed method, unlike [4], can automatically pick only one plane rather than many planes in [4]. Moreover, in contrast to [21], it does not require high learning time.

Table 1: Apex spotting accuracy (%) on CASME with 15 training samples

Methods	Acc.
LBP on the 3 orthogonal planes [16]	31.2%
Cubic-LBP [2]	38.0%
LBP [17]	39.0%
LBP on the 4 planes [3]	45.1%
Combining PDE and Cubic-ULBP (ours)	**81.2%**

Table 2: Micro-changes spotting accuracy (%) on CASME II with 15 training samples

Methods	Acc.
LBP [17]	10.3%
Cubic-LBP [2]	20.1%
Combining PDE and Cubic-ULBP (ours)	**61.0%**

Table 3: Micro-changes spotting accuracy (%) using the proposed method (Combining PDE and Cubic-ULBP)

Dataset	Acc.
POVI	83.3%
FA	85.1%

Table 4: The obtained mean absolute error and standard error in our work

Dataset	mean absolute error	standard error
POVI	0.72	0.045
CASME	0.80	0.050
FA	0.63	0.040
CASME II	4.46	0.52

Table 5: Productivity of our proposed method (Combining PDE and Cubic-ULBP) using other metrics

Dataset	recall	$prec.$	F1
FA	0.82	0.86	0.84
CASME	0.79	0.82	0.81
POVI	0.84	0.81	0.82
CASME II	0.59	0.61	0.60

Table 6: Spotting accuracy (%) on CASME, CASME II, POVI, and FA with 5, 10, and 15 training samples

Method	Dataset	# training samples		
		5	10	15
Combining PDE and Cubic-ULBP (ours)	CASME	76.4	80.3	81.2
	CASME II	54.3	59.4	61.0
	POVI	77.8	82.1	83.3
	FA	79.1	83.9	85.1

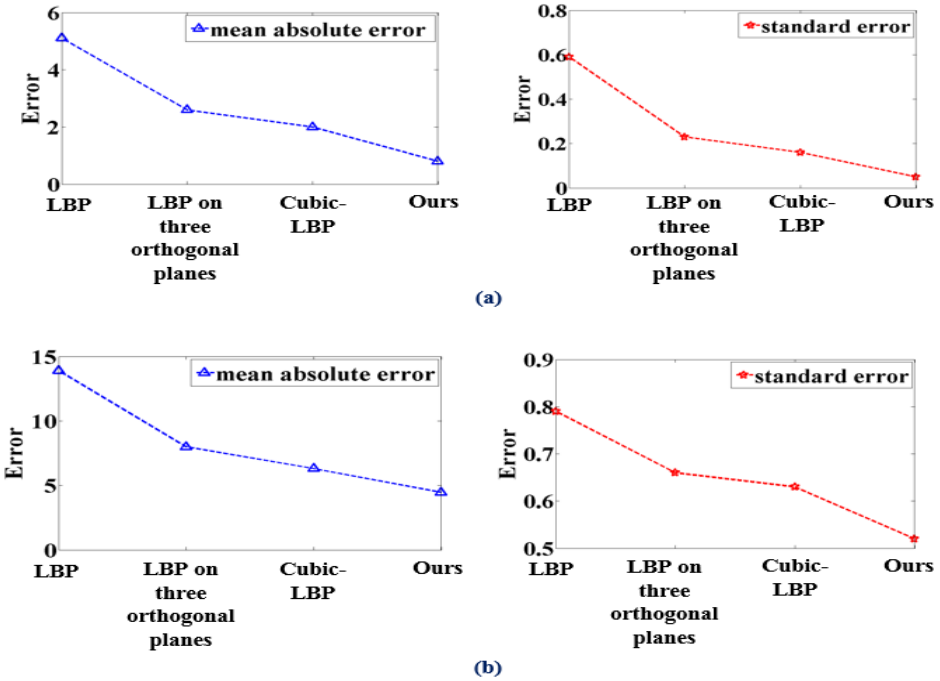


Fig. 7: Comparing error in other related methods and ours. (a) CASME (b) CASME II.

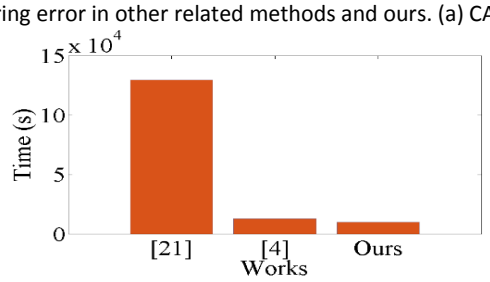


Fig. 8: Comparing the elapsed time of our proposed method with the existing works.

Conclusion

In our planet and even in cosmos, there are plenty of micro-changes. They contain significant information which should not be underestimated. Thus, spotting such tiny movements could be a bottleneck in wide range of applications from recognizing a liar to the health care system. Nevertheless, micro-change characteristics, namely low-intensity and brief appearance, make them almost invisible. In order to cope with this challenge, we proposed the novel approach which takes advantage of both Cubic-ULBP and PDE. Cubic-ULBP is robust against noise and illumination variation and at the same time could spot micro-changes in any direction. PDE is then applied to extract most effective features. According to our results, micro-change within the apex frame is located with a satisfying accuracy. In the future, the proposed PDE on the Cubic-ULBP could be used in similar scenarios, where the micro-changes need to be spotted. The limitation of the current work is the scarce number of 3D public datasets.

Author Contributions

Vida Esmaeili: Conceptualization, Methodology, Investigation, Visualization, Analysis and interpretation, Data curation, Writing - original draft.

Mahmood Mohassel Feghhi: Methodology, Investigation, Writing - review & editing, Supervision, Data curation, Validation, Project administration, Formal analysis.

Seyed Omid Shahdi: Conceptualization, Investigation, Data curation, Visualization, Supervision, Validation, Analysis and interpretation, Writing - review & editing.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

PDE	Partial Differential Equations
Cubic-ULBP	Cubic Uniform Local Binary Pattern
LBP-TOP	Local Binary Pattern on Three Orthogonal Planes
FA	Fluorescein Angiography

References

[1] P. Ekman, W. V. Friesen, "Nonverbal leakage and clues to deception," Psychiatry, 32(1): 88-106, 1969.

[2] B. Yang, J. Cheng, Y. Yang, B. Zhang, J. Li, "MERTA: micro-expression recognition with ternary attentions," Multimedia Tools Appl., 80(11): 1-16, 2021.

[3] L. Zhou, X. Shao, Q. Mao, "A survey of micro-expression recognition," Image Vision Comput., 105: 104043, 2021.

[4] V. Esmaeili, S. O. Shahdi, "Automatic micro-expression apex spotting using Cubic-LBP," Multimedia Tools Appl., 79(27): 20221-20239, 2020.

[5] V. Esmaeili, M. M. Feghhi, S. O. Shahdi, "Micro-Expression recognition using histogram of image gradient orientation on diagonal planes," in Proc. 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA): 1-5, 2021.

[6] V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, "Autonomous apex detection and micro-expression recognition using proposed diagonal planes," Int. J. Nonlinear Anal. Appl., 11: 483-497, 2020.

[7] V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, "A comprehensive survey on facial micro-expression: approaches and databases," Multimedia Tools Appl., 81: 40089- 40134, 2022.

[8] V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, "Automatic micro-expression apex frame spotting using local binary pattern from six intersection planes," arXiv preprint arXiv: 2104.02149, 2021.

[9] M. Xiao, Y. Chen, J. Zhu, H. Zhang, X. Zhao, L. Gao, W. Xing, "Climbing the apex of the ORR volcano plot via binuclear site construction: electronic and geometric engineering," J. the Am. Chem. Soc., 141(44): 17763-17770, 2019.

[10] K. S. Exner, "Does a thermoneutral electrocatalyst correspond to the apex of a volcano plot for a simple two-electron process?," Angew. Chem. Int. Ed., 59(26): 10236-10240, 2020.

[11] C. Widiwijayanti, J. Déverchère, R. Louat, M. Sébrier, H. Harjono, M. Diament, D. Hidayat, "Aftershock sequence of the 1994, Mw 6.8, Liwa earthquake (Indonesia): Seismic rupture process in a volcanic arc," Geophys. Res. Lett., 23(21): 3051-3054, 1996.

[12] J. Skinner, "The smoke of an eruption and the dust of an earthquake: Dark tourism, the sublime, and the re-animation of the disaster location," in the Palgrave handbook of dark tourism studies. Palgrave Macmillan, London: 125-150, 2018.

[13] O. Yildirim, M. Talo, E. J. Ciaccio, R. San Tan, U. R. Acharya, "Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ECG records," Comput. Methods Programs Biomed., 197: 105740, 2020.

[14] E. Moghadas, J. Rezazadeh, R. Farahbakhsh, "An IoT patient monitoring based on fog computing and data mining: Cardiac arrhythmia usecase," Internet Things, 11: 100251, 2020.

[15] S. T. Liong, J. See, K. Wong, R. C. W. Phan, "Less is more: Micro-expression recognition from video using apex frame," Signal Process. Image Commun., 62: 82-92, 2018.

[16] Y. Li, X. Huang, G. Zhao, "Can micro-expression be recognized based on single apex frame?," in Proc. 2018 25th IEEE International Conference on Image Processing (ICIP): 3094-3098, 2018.

[17] G. Zhao, M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," IEEE Trans. Pattern Anal. Mach. Intell., 29(6): 915-928, 2007.

[18] T. Ojala, M. Pietikäinen, D. Harwood, "A comparative study of texture measures with classification based on featured distributions," Pattern Recogn., 29(1): 51-59, 1996.

[19] T. Ojala, M. Pietikainen, T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Anal. Mach. Intell., 24(7): 971-987, 2002.

[20] V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, "Micro-expression recognition based on the multi-color ulbp and histogram of gradient direction from six intersection planes," J. Iran. Associ. Electr. Electron. Eng., 19(3): 123-130, 2022.

[21] V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, "Spotting micro-movements in image sequence by introducing intelligent cubic-LBP," IET Image Process., 16(14): 3814-3830, 2022.

[22] V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, "Early COVID-19 diagnosis from lung ultrasound images combining RIULBP-TP and 3D-DenseNet," in Proc. 2022 9th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS): 1-5, 2022.

[23] V. Esmaeili, M. Mohassel Feghhi, "Real-time authentication for electronic service applicants using a method based on two-stream 3d deep learning," Soft Comput. J., 2023.

[24] V. Esmaeili, M. Mohassel Feghhi, "Diagnosis of Covid-19 disease by combining hand-crafted and deep-learning methods on ultrasound data," J. Mach. Vision Image Process., 9(4): 31-41, 2022.

[25] V. Esmaeili, M. Mohassel Feghhi, "COVID-19 diagnosis: ULBPFP-net approach on lung ultrasound data," Iran. J. Electr. Electron. Eng., 19(3): 2586-2586, 2023.

[26] V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, "Automatic micro-expression recognition using LBP-SIPl and FR-CNN," AUT J. Model. Simul., 54(1): 59-72, 2022.

[27] N. Wadhwa, M. Rubinstein, F. Durand, W. T. Freeman, "Phase-based video motion processing," ACM Trans. Graphics (TOG), 32(4): 1-10, 2013.

[28] H. Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," ACM Trans. Graphics (TOG), 31(4): 1-8, 2012.

[29] G. Fahmy, M. F. Fahmy, O. M. Fahmy, "Micro-movement magnification in video signals using complex wavelet analysis," IET Image Process., 11(11): 986-993, 2017.

[30] C. Liu, A. Torralba, W. T. Freeman, F. Durand, E. H. Adelson, "Motion magnification," ACM Trans. Graphics (TOG), 24(3): 519-526, 2005.

[31] M. Janatka, A. Sridhar, J. Kelly, D. Stoyanov, "Higher order of motion magnification for vessel localisation in surgical video," in Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention: 307-314, 2018.

[32] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," IEEE Trans. Affective Comput., 9(4): 563-577, 2017.

[33] M. Lu, Y. Chai, Q. Liu, "Observation of tower vibration based on subtle motion magnification," IFAC-PapersOnLine, 52(24): 346-350, 2019.

[34] W. Fan, Z. Zheng, W. Zeng, Y. Chen, H. Q. Zeng, H. Shi, X. Luo, "Robotically surgical vessel localization using robust hybrid video motion magnification," IEEE Rob. Autom. Lett., 6(2): 1567-1573, 2021.

[35] A. Mohseni, "A new PDE-based resolution enhancement technique for the analysis of low SNR particle displacement images," Eur. J. Mech. B.Fluids, 85: 289-311, 2021.

[36] T. Barbu, "Feature keypoint-based image compression technique using a well-posed nonlinear fourth-order PDE-based model," Mathematics, 8(6): 930, 2020.

[37] L. Afraites, A. Hadri, A. Laghrib, M. Nachaoui, "A high order PDE-constrained optimization for the image denoising problem," Inverse Probl. Sci. Eng., 29(12):1821-1863, , 2021.

[38] A. K. Al-Jaberi, E. M. Hameed, "Topological data analysis for evaluating PDE-based denoising models," J. Phys. Conf. Ser., 1897(1): 012006. 2021.

[39] C. Fang, Z. Zhao, P. Zhou, Z. Lin, "Feature learning via partial differential equation with applications to face recognition," Pattern Recogn., 69: 14-25, 2017.

[40] S. Moorthi, S. Karthikeyan, "A Study On Learning Partial Differential Equations Algorithm For Face Recognition," Int. J. Innovative Res. Explorer, 5(4): 127-134, 2018.

[41] W. J. Yan, Q. Wu, Y. J. Liu, S. J. Wang, X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in Proc. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG): 1-7, 2013.

[42] W. J. Yan, X. Li, S. J. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," PloS one, 9(1): e86041, 2014.

[43] A. J. Witsil, J. B. Johnson, "Volcano video data characterized and classified using computer vision and machine learning algorithms," Geosci. Front., 11(5): 1789-1803, 2020.

[44] W. Fuhl, T. Kübler, K. Sippel, W. Rosenstiel, E. Kasneci, "Excuse: Robust pupil detection in real-world scenarios," in Proc. International Conference on Computer Analysis of Images and Patterns: 39-51, 2015.

[45] C. Nicholson, "Evaluation metrics for machine learning—accuracy, precision, recall, and F1 defined," ed: Pathmind.

Biographies

Vida Esmaeili received her B.S. degree in Electrical Engineering from the Azad University of Abhar, Iran, in 2015 and the M.S. degree in Electrical Engineering from the Azad University of Qazvin, Iran, in 2018. Since 2019, she has been working toward the Ph.D. degree in Electrical Engineering at the Tabriz University, Iran. Her research interests include the area of image processing, machine learning, and pattern recognition.

- Email: v.esmaeili@tabrizu.ac.ir
- ORCID: 0000-0002-1840-8659
- Web of Science Researcher ID: AAB-9907-2022
- Scopus Author ID: 57215699025
- Homepage: https://scholar.google.com/citations?user=lZvaGUYAAAAJ&hl=en

Mahmood Mohassel Feghhi received the B.S. and M.S. degrees (with Honors) in Electrical Engineering from the Iran University of Science and Technology, Tehran, Iran, in 2006 and 2009, respectively, and the Ph.D. degree in Electrical Engineering from the College of Engineering, University of Tehran, Iran, in 2015. From 2007 to 2016, he held positions as a senior design engineer in the areas of communication systems design with several Communications Industries and Inc. Since 2016, he has been with the Faculty of Electrical and Computer Engineering, University of Tabriz, Iran, where he is now an associate professor. His current research interests include network information theory, wireless networks, machine learning, resource allocation, scheduling and optimization.

- Email: mohasselfeghhi@tabrizu.ac.ir
- ORCID: 0000-0002-7193-843X
- Web of Science Researcher ID: D-2414-2010
- Scopus Author ID: 27267677800
- Homepage: https://asatid.tabrizu.ac.ir/en/pages/default.aspx?mohasselfeghhi

Seyed Omid Shahdi received his B.S. degree in Electrical Engineering from the Azad University of Yazd, Iran, in 2006 and the M.S. degree in Electrical Engineering from the Azad University of Qazvin, Iran, in 2009 and the Ph.D. degree from the Universiti Teknologi Malaysia, Johor (Malaysia), in 2012. Currently, he is an assistant professor in the faculty of Electrical, Biomedical and Mechatronics Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran. His research interests include the areas of machine learning, pattern recognition, neural networks, and image processing. He has published many papers both at the national and international levels.

- Email: shahdi@qiau.ac.ir
- ORCID: 0000-0002-1827-1883
- Web of Science Researcher ID: AAV-8113-2021
- Scopus Author ID: 36680945800
- Homepage: https://old.qazvin.iau.ir/teacher/teacher.aspx?number=913780

How to cite this paper:
V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, "Applying partial differential equations on cubic uniform local binary pattern to reveal micro-changes," J. Electr. Comput. Eng. Innovations, 12(1): 259-270, 2024.

DOI: 10.22061/jecei.2023.9600.639

URL: https://jecei.sru.ac.ir/article_2012.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Estimation of Wheel-Rail Adhesion Force Using Traction System Behavior

M. Moradi, R. Havangi *

Faculty of Electrical Engineering and Computer, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 14 August 2023
Reviewed 17 October 2023
Revised 14 November 2023
Accepted 04 December 2023

Keywords:
Extended Kalman Filter
Adhesion model
Railway traction
Torque
Estimation

Corresponding Author's Email Address:
Havangi@Birjand.ac.ir

Abstract

Background and Objectives: Traction system and adhesion between wheel and rail are fundamental aspects in rail transportation. Depending on the vehicle's running conditions, different levels of adhesion are needed. Low adhesion between wheel and rail can be caused by leaves on the line or other contaminants, such as rust or grease. Low adhesion can occur at any time of year especially in autumn, resulting in disruptions to passenger journeys. Increased wheel-rail adhesion for transit rail services results in better operating performance and system cost savings. Deceleration caused by low adhesion, will extend the braking distance, which is a safety issue. Because of many uncertain or even unknown factors, adhesion modelling is a time taking task. Furthermore, as direct measurement of adhesion force poses inherent challenges, state observers emerge as the most viable choice for employing indirect estimation techniques. Certain level of adhesion between wheel and rail leads to reliable, efficient, and economical operation.
Methods: This study introduces an advantageous approach that leverages the behavior of traction motors to provide support in achieving control over wheel slip and adhesion in railway applications. The proposed method aims to enhance the utilization of existing adhesion, minimize wheel deterioration, and mitigate high creep levels. In this regard, estimation of wheel-rail adhesion force is done indirectly by concentrating on induction motor parameters as railway traction system and dynamic relationships. Meanwhile, in this study, we focus on developing and applying the sixth-order Extended Kalman Filter (EKF) to create a highly efficient sensorless re-adhesion control system for railway vehicles.
Results: EKF based design is compared with Unscented Kalman Filter (UKF) based and actual conditions and implemented in Matlab to check the accuracy and performance ability for state and parameter estimation. Experimental results showed fast convergence, high precision and low error value for EKF.
Conclusion: The proposed technique has the capability to identify and assess the current state of local adhesion, while also providing real-time predictions of wear. Besides, in combination with control methods, this approach can be very useful in achieving high wheel-rail adhesion performance under variable complex road conditions.

Introduction

The contact force at wheel and rail interface governs the dynamic behavior of entire vehicle, which is complex and highly non-linear. Measurement of this force is one of the most important issues for condition monitoring and safety evaluation of railway vehicles. In [1] an estimator framework is presented for online identification of contact force at wheel- rail interface. Sliding and slipping are two challenging situations in railway industry that arise from the low friction between wheel and rail,

especially when wheel and rail are contaminated by different factors such as mud, grease, humidity, etc [2]-[5]. Also, weather conditions [6], deliberately applied friction modifiers [7], or contact surface temperatures [8]-[10], can affect the amount of adhesion. In order to avoid wheel slide/slip and uncomfortable riding and decrease in traction effort, wheel wear, and noise, it is imperative to minimize the excessive slippage that occurs between the surfaces of the wheel and the rail. In [11], the proposed method to reveal the slip is to compare speed difference between the wheel and the vehicle body. Then the estimated slip is used for torque compensation signal generation. Since the induction motor is one of the most important parts of the train's motion system, investigation of the induction motor was proposed which uses the estimated adhesion force to suppress the slip and adjust the torque command [12].

Many researchers tried to resolve adhesion problem and different solutions such as mathematical control theory, statistical and genetic have been proposed and applied [13], [14]. The adhesion characteristic has two stable and unstable areas. Between these two areas, maximum value of the adhesion is located. The adhesion coefficient depends upon the slip velocity, which influences on adhesion coefficient level. Train velocity and temperature of contact area are two important factors affecting the railway surfaces. Higher values of the adhesion coefficient and the slip velocity lead to maximum adhesion coefficient. Therefore, adhesion level identification is an important task for proper operation of a railway vehicle. A novel approach was introduced in a recent study [15] to determine the adhesion coefficient between the wheel and rail. Additionally, another research paper [16] presents a distinct adhesion control technique that relies on observing the adhesion state between the wheel and rail. Obtaining optimal adhesion control can lead to effective utilization of train traction power [17], [18]. It is important to mention that based on the changes observed in the characteristic curve of the adhesion coefficient, it is necessary to limit the creep velocity of the train within the stable region to prevent wheel slide or slip. To bring the trains back to the stable region, readhesion control is implemented by finely tuning the torque and promptly detecting instances of wheel slide or slip. However, a limitation of this approach is that it is unable to completely eliminate the occurrence of slide or slip [19]. To explore the phenomenon of slide and slip in railway traction, a novel approach utilizing the second-order Luenberger observer is introduced. This method indirectly determines the frictional force associated with this phenomenon [20]. A bank of Kalman Filter (KF) is applied for the adhesion estimation.

Identification of the contact conditions is then done by examining the residuals from the Kalman filters [21].

In [22] a Kalman Filter based technique is proposed for estimation of low adhesion between wheel and rail. The EKF is the nonlinear form of Kalman Filter, which has been used extensively for estimation of nonlinear states in navigation systems [23]. Extended Kalman filter based estimation for estimating the creepage, creep force, and friction coefficient between the wheel and rail surfaces by utilizing the AC motor parameters such as stator voltage, current, and speed was proposed in [24]. An alternative approach to detect slip velocity is through the utilization of multi-rate Extended Kalman Filter state identification. This method combines both the multi-rate technique and the EKF method to accurately determine the traction motor load torque. The advantages of this method are faster slip detection and improved reliability and traction performance [25]. To predict the wheel and rail wear, regions of adhesion variations or low adhesion, and the development of rolling contact fatigue, a novel approach utilizing the Kalman-Bucy filter technique is suggested to estimate the wheel and the rail states [26]. In [27], a model-based technique is proposed for condition monitoring, in which an unscented Kalman filter is applied to estimate rolling radius by considering the angular velocity and the traction effort of the motor measurements. In [28]-[30] UKF was used for sensorless speed control of induction motor, in which it was emphasized that UKF has more robust estimation performance.

This study investigates the utilization of EKF approach to accurately determine the adhesion force between the contact surfaces of a wheel and rail. The estimation is achieved through analyzing the measured values of the stator currents of an induction motor. To evaluate the observer's performance, a dynamic model is constructed, comprising a gear box, wheelset, and induction motor. The behavior of the wheel-rail contact is described using the Polach model. The design of the induction motor is based on a first-order decomposition of the sixth-order nonlinear model. The proposed EKF technique is capable of estimating various parameters, including motor current, rotor flux components, motor speed, and load torque.Then, dynamic reletion is used for adhesion force estimation. For further investigation, we compared our method with UKF. The obtained results show good convergence and high precision. The rest of this research is organized into four parts. First, the details of traction system and mathematical model of induction motor are explained. Then, the estimator framework is presented. Following this, the details of experimental results are highlighted. Finally, the conclusion is given.

System Modelling and Discretization

The traction system information employed in this research is shown in Fig. 1. The model consists of three parts, wheel and rail, gear wheel, and traction motor. The

continuous dynamic model of the induction motor used in this research is described by sixth-order nonlinear differential equations with three series of variables consisting of two mechanical variables (motor speed and load torque), four electrical variables (currents and fluxes), and two control variables (stator voltages) and the stationary reference frame is (α, β). The action of the axle load causes the wheels rotation, which leads to micro deformation region occurring in the wheel-rail contact region. Then, the interaction between wheel and rail produces the adhesion force F_a. The schematic of wheel-rail adhesion mechanism is shown in Fig. 2. The states, the measurements, the stator voltages, and the state and measurement noises are given in (1) to (5) respectively.

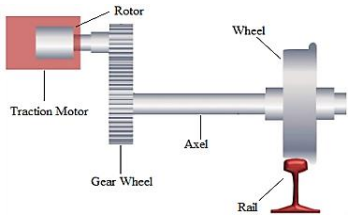


Fig. 1: Schematic of the traction system.

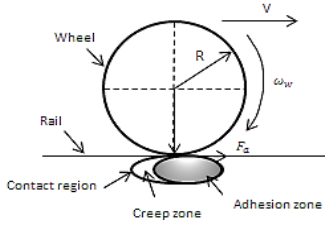


Fig. 2: Wheel-rail adhesion mechanism.

$$x(t) = [x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6]^T = [I_{s\alpha} \quad I_{s\beta} \quad \psi_{r\alpha} \quad \psi_{r\beta} \quad \omega_m \quad T_L]^T \tag{1}$$

$$z = [\, I_{s\alpha} \quad I_{s\beta}]^T \tag{2}$$

$$u(t) = [\, u_{s\alpha} \quad u_{s\beta}]^T \tag{3}$$

$$w(t) = [w_1 \quad w_2 \quad w_3 \quad w_4 \quad w_5 \quad w_6]^T \tag{4}$$

$$v = [v_1 \quad v_2]^T \tag{5}$$

where $I_{s\alpha}$ is stator current in α frame, $I_{s\beta}$ is stator current in β frame, $\psi_{r\alpha}$ is rotor flux in α frame, $\psi_{r\beta}$ is rotor flux in β frame, ω_m is the motor angular velocity, and T_L is load torque.

The equations are listed as follows [31]:

$$\frac{dI_{S\alpha}}{dt} = -\left(\frac{R_s}{\sigma L_s} + \frac{L_m^2 R_r}{\sigma L_s L_r^2}\right) I_{s\alpha} + \frac{L_m R_r}{\sigma L_s L_r^2}\psi_{r\alpha} + \frac{L_m}{\sigma L_s L_r} n_p \omega_m \psi_{r\beta} + \frac{1}{\sigma L_s} u_{s\alpha} \tag{6}$$

$$\frac{dI_{S\beta}}{dt} = -\left(\frac{R_s}{\sigma L_s} + \frac{L_m^2 R_r}{\sigma L_s L_r^2}\right) I_{s\beta} - \frac{L_m}{\sigma L_s L_r} n_p \omega_m \psi_{r\alpha} + \frac{L_m R_r}{\sigma L_s L_r^2}\psi_{r\beta} + \frac{1}{\sigma L_s} u_{s\beta} \tag{7}$$

$$\frac{d\psi_{r\alpha}}{dt} = \frac{R_r L_m}{L_r} I_{s\alpha} - \frac{R_r}{L_r}\psi_{r\alpha} - n_p \omega_m \psi_{r\beta} \tag{8}$$

$$\frac{d\psi_{r\beta}}{dt} = \frac{R_r L_m}{L_r} I_{s\beta} + n_p \omega_m \psi_{r\alpha} - \frac{R_r}{L_r}\psi_{r\beta} \tag{9}$$

$$\frac{d\omega_m}{dt} = \frac{-3 n_p L_m}{2 J_{eqv} L_r}\psi_{r\beta} I_{s\alpha} + \frac{3 n_p L_m}{2 J_{eqv} L_r}\psi_{r\alpha} I_{s\beta} - \frac{C_v}{J_{eqv}}\omega_m - \frac{1}{J_{eqv}} T_L \tag{10}$$

$$\frac{dT_L}{dt} = 1 \tag{11}$$

where R_s is the stator resistance, R_r is the rotor resistance, L_s is the stator self-inductance, L_r the is rotor self-inductance, L_m is the mutual inductance, n_p is the number of the pole pairs, J_{eqv} is the equivalent moment of inertia, C_v is the viscous friction, and σ is the leakage coefficient and defined as (12).

$$\sigma = 1 - \frac{L_m^2}{L_s L_r} \tag{12}$$

Induction motor extended model is shown in (13).

$$\underbrace{\begin{bmatrix} \dot{I}_{s\alpha} \\ \dot{I}_{s\beta} \\ \dot{\psi}_{r\alpha} \\ \dot{\psi}_{r\beta} \\ \dot{\omega}_m \\ \dot{T}_L \end{bmatrix}}_{\dot{x}_e} = \underbrace{\begin{bmatrix} -\left(\frac{R_s}{\sigma L_s} + \frac{L_m^2 R_r}{\sigma L_s L_r^2}\right) & 0 & \frac{L_m R_r}{\sigma L_s L_r^2} & \frac{L_m}{\sigma L_s L_r} n_p \omega_m & 0 & 0 \\ 0 & -\left(\frac{R_s}{\sigma L_s} + \frac{L_m^2 R_r}{\sigma L_s L_r^2}\right) & -\frac{L_m}{\sigma L_s L_r} n_p \omega_m & \frac{L_m R_r}{\sigma L_s L_r^2} & 0 & 0 \\ \frac{R_r L_m}{L_r} & 0 & -\frac{R_r}{L_r} & -n_p \omega_m & 0 & 0 \\ 0 & \frac{R_r L_m}{L_r} & n_p \omega_m & -\frac{R_r}{L_r} & 0 & 0 \\ \frac{-3 n_p L_m}{2 J_{eqv} L_r}\psi_{r\beta} & \frac{3 n_p L_m}{2 J_{eqv} L_r}\psi_{r\alpha} & 0 & 0 & -\frac{C_v}{J_{eqv}} & -\frac{1}{J_{eqv}} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{A_e} \underbrace{\begin{bmatrix} I_{s\alpha} \\ I_{s\beta} \\ \psi_{r\alpha} \\ \psi_{r\beta} \\ \omega_m \\ T_L \end{bmatrix}}_{x_e} + \underbrace{\begin{bmatrix} \frac{1}{\sigma L_s} & 0 \\ 0 & \frac{1}{\sigma L_s} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}}_{B_e} \underbrace{\begin{bmatrix} U_{s\alpha} \\ U_{s\beta} \end{bmatrix}}_{U_e} + w(t) \tag{13}$$

$$\underbrace{\begin{bmatrix} I_{s\alpha} \\ I_{s\beta} \end{bmatrix}}_{z} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}}_{H_e} \begin{bmatrix} I_{s\alpha} \\ I_{s\beta} \\ \psi_{r\alpha} \\ \psi_{r\beta} \\ \omega_m \\ T_L \end{bmatrix} + v(t) \tag{14}$$

Ignoring the damping coefficient, the dynamic equation of traction motor is as follows [13]:

$$\frac{d\omega_m}{dt} = \frac{T_m - T_L}{J_{eqv}} \tag{15}$$

$$\omega_w = \frac{\omega_m}{n_i} \tag{16}$$

$$T_m = \frac{n_p L_m}{L_r}(I_{s\beta}\psi_{r\alpha} - I_{s\alpha}\psi_{r\beta}) \quad (17)$$

$$T_L = \frac{2rF_a}{n_i} \quad (18)$$

$$J_{eqv} = J_m + \frac{J_g + J_x + J_{wR} + J_{wL}}{n_i^2} \quad (19)$$

where $\omega_w$represents the angular velocity of the wheel, while Fa denotes the adhesion force exerted by a single wheel. Additionally, J_m, J_g, J_x, J_{wR}, and J_{wL} refer to the moments of inertia associated with the motor, gearbox, wheelset axle, right wheel, and left wheel, respectively. The adhesion force at the contact point between the wheel and rail, denoted as F_a is determined using Polach's method [33] and can be calculated using (20).

$$F_a = \frac{2F_N\mu_f}{\pi}\left(\frac{k_A\epsilon}{1+(k_A\epsilon)^2} + \arctan(k_S\epsilon)\right),\ k_S \le k_A \le 1 \quad (20)$$

where F_N is the normal force between the wheel and rail, μ_f is the traction coefficient, and quantities k_A and k_S are Polach reduction factors in the areas of adhesion and slip, respectively.

$$\mu_f = \mu_0((1-D)e^{-B\xi V} + D) \quad (21)$$

$$\epsilon = \frac{G\pi abC_{11}}{4F_N\mu_f}\xi \quad (22)$$

where *D* and *B* represent reduction factors associated with distinct friction coefficients, V is the longitudinal velocity of the train, ξ is the creepage between the wheel and rail, G is shear module, a and b are the semi-axis length of the contact patch and C_{11} is the Kalker coefficient.

The creepage contains longitudinal and lateral components but in this research, the lateral dynamics are neglected, so calculated by the following equation [34]:

$$\xi = \frac{\omega_w r - V}{V} \quad (23)$$

Estimation of Wheel-Rail Adhesion

The details of the EKF and UKF used for estiomation of Wheel-Rail adhesion can be found in the following subsections.

A. Extended Kalman Filter

The EKF is an enhanced variant of the traditional Kalman filter that takes into account nonlinear systems. In this study, our goal is to determine the optimal linear estimation for the state vector of the induction motor. The discrete-time nonlinear model is expressed as below:

$$x_{k+1} = f(x_k, u_k, w_k) \quad (24)$$

$$z_k = h(x_k, v_k) \quad (25)$$

where $f(.)$ represents the dynamics of machine, h(.) is the relationship between the observation z_k and the state vector x_k, $u_k$refers to the input provided to the motor, while w_k and v_k represent the vectors of noise that affect the process and measurement respectively. Equations (24) and (25) exhibit nonlinearity, necessitating their linearization. This process involves employing the first-order Taylor approximation in the vicinity of a chosen reference point. Linearizing these nonlinear equations will result in the following description of the dynamics:

$$x_{k+1} = f(\hat{x}_k, u_k, 0) + F_K(x_k - \hat{x}_k) + W_k w_k \quad (26)$$

$$z_k = h(\hat{x}_k, 0) + H_K(x_k - \hat{x}_k) + V_k(v_k - 0) \quad (27)$$

where F_K, W_k, H_K and V_k are Jacobean matrices defined as below:

$$F_K = \frac{\partial f}{\partial x}\Big|_{\hat{x}_k}, W_k = \frac{\partial f}{\partial w}\Big|_{w}, H_k = \frac{\partial h}{\partial x}\Big|_{\hat{x}_k}, V_k = \frac{\partial h}{\partial v}\Big|_{\hat{x}_k} \quad (28)$$

The EKF algorithm using induction motor model in (13) and (14) can be given by the following equations:

$$P_{k+1|k} = F_K P_k(k) {F_K}^T + W_k Q {W_K}^T \quad (29)$$

$$K_K = P_{k+1|k}{H_k}^T(H_k P_{k+1|k}{H_k}^T + V_k R {V_K}^T)^{-1} \quad (30)$$

$$\hat{x}_{k+1|k} = f(\hat{x}_{k|k}.u_k, 0) \quad (31)$$

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_K(z_k - h(\hat{x}_{k+1|k}, 0)) \quad (32)$$

$$P_{k+1|k+1} = (I - K_K H_k)P_{k+1|k} \quad (33)$$

where $P_{k+1|k}$ is the priori prediction error covariance matrix, $P_{k+1|k+1}$ is the posteriori prediction error covariance matrix, K_K is the Kalman gain, $\hat{x}_{k+1|k}$ is the priori state prediction vector, $\hat{x}_{k+1|k+1}$is the posteriori state prediction vector and Q and R are the covariance matrixes of process and measurement noise and I is the unit matrix symbol. In general, the extended Kalman filter is not an optimal estimator. If the process is modeled incorrectly, or if the initial estimate of the state is wrong, linearization may lead to rapid divergence of the filter. Furthermore, the estimated covariance matrix in EKF has a tendency to inaccurately assess the true covariance matrix. Consequently, it runs the risk of losing consistency in the statistical context unless stabilizing noise is introduced. Finally, because of the Q and R uncertainty, their values are obtained by trial-and-error methods which is tedious and time-consuming procedure.

B. Unscented Kalman Filter

The UKF is created by incorporating the unscented transformation (UT) method.It is assumed that the studied system is nonlinear and in discrete form:

$$x_k = f(\hat{x}_k, u_k) + w_k \qquad w_k \sim (0, Q_k) \quad (34)$$

$$z_k = h(\hat{x}_k, u_k) + v_k \qquad v_k \sim (0, R_k) \quad (35)$$

In the first step of estimate of state vector of the induction motor using UKF, a set of $2n_x + 1$ weighted samples or sigma points are determined as:

$$\chi_{k-1} = [\hat{x}_{k-1} \quad \hat{x}_{k-1} + \sqrt{(n_x + \lambda)(P_{k-1})_i} \quad \hat{x}_{k-1} - \sqrt{(n_x + \lambda)(P_{k-1})_i}] \quad (36)$$

$$w_m^{(0)} = \frac{\lambda}{\lambda+n_x} \quad (37)$$

$$w_c^{(0)} = \frac{\lambda}{\lambda+n_x} + 1 - \alpha^2 + \beta \quad (38)$$

$$w_c^{(j)} = w_m^{(j)} = \frac{\lambda}{2(\lambda+n_x)} \quad (39)$$

where the dimension of the state variable is represented as n_x .The estimate of $x_k$at time k-1 is denoted as $\hat{x}_{k-1}$, and its covariance is represented as P_{k-1}. The weight $w_m$is is utilized for determining the mean, while w_c is employed for calculating the covariance. The parameter α, which lies within the range of [0,1], is employed to regulate the distribution of the sigma points. Additionally, β non-negative term, is utilized to incorporate information from higher order moments of the distribution and $\lambda = \alpha^2(n_x + \rho) - n_x$. It should be noted that in this study, these three parameters are set as follows: $\alpha = 1$, $\beta = 0$ and $\rho = 1$

The column i of the matrix $P_{k-1}$is denoted as $(P_{k-1})_i$. Sigma points χ_{k-1} are substituted into the nonlinear state equation, and the transformed sigma points are evaluated for each of the the 0 - $2n_x$ points as described below:

$$\chi_k^{(i)} = f(\chi_{k-1}^{(i)}, u_k) \quad (40)$$

To obtain the mean and covariance of the modified set of sigma points, the following procedure is employed:

$$\hat{x}_k^- = \sum_{i=1}^{2n_x} w_m^{(i)} \chi_k^{(i)} \quad (41)$$

$$P_k^- = \sum_{i=1}^{2n_x} w_c^{(i)} (\chi_k^{(i)} - \hat{x}_k^-)(\chi_k^{(i)} - \hat{x}_k^-)^T + Q_k \quad (42)$$

where Q_k is the process noise covariance. The sigma points that have been transformed are subsequently utilized to predict the measurements by employing the measurement model:

$$\xi^{(i)} = h(\chi_k^{(i)}, U_k) \quad (43)$$

The expected measurement $\hat{z}_k$is as:

$$\hat{z}_k = \sum_{i=1}^{2n_x} w_m^{(i)} \xi^{(i)} \quad (44)$$

Using the predicted sigma points, $P_k^{xz}$and P_k^{zz} also determines as follows:

$$P_k^{xz} = \sum_{i=0}^{2n} \omega_i^{(c)} (\chi_k^{(i)} - \hat{x}_k)(\xi^{(i)} - \hat{z}_k^-)^T \quad (45)$$

$$P_k^{zz} = \sum_{i=0}^{2n} \omega_i^{(c)} (\xi^{(i)} - \hat{z}_k^-)(\xi^{(i)} - \hat{z}_k^-)^T + R_k \quad (46)$$

The mean and square root of covariance for the state are recalculated based on the actual measurement.

$$\hat{x}_k = \hat{x}_k + K_k(z_k - \hat{z}_k) \quad (47)$$

$$P_k = P_k^- - K_K P_k^{xz} {K_K}^T \quad (48)$$

$$K_k = P_k^{xz}(P_k^{zz})^{-1} \quad (49)$$

Results

This section begins by simulating the presented model to verify the accuracy of the EKF in estimating variables. Subsequently, the performance of the EKF is assessed by comparing it with the UKF to determine its accuracy as an estimator. All of our codes have been developed and implemented using the Matlab, with a sampling period of 10^{-3}s seconds. To ensure more realistic testing conditions, the induction motor is powered through an AC drive with a sinusoidal input voltage.

A. EKF-Based Model Simulation

In the first step of our simulation, we try to simulate contact conditions. Our goal in this step is to show the created changes in adhesion force versus creepage for all track conditions such as dry, wet, low, and very low relationship between adhesion force and creepage. The designed friction coefficients are as follows:

$$\mu_0 = \begin{cases} 0.55 & t < 10 \\ 0.3 & 10 \le t < 20 \\ 0.06 & 20 \le t < 30 \\ 0.03 & 30 \le t < 35 \end{cases}$$

The values of k_A, k_S, *D* and *B* under different friction conditions are listed in Table 1 and the other parameter values used in equations (20) to (23) are as the following: F_N = 60 KN, *G* = 8.4×10^{10} $\frac{N}{m^2}$, *a* = 0.0015 m, *b* = 0.0075 m, C_{11} =4.12, *V*= 15 $\frac{m}{s}$

Fig. 3 shows the curves of the adhesion force versus creepage in different wheel-rail contact conditions. As the creepage increases, the slip region increases versus the stick region. As we see, the adhesion force changeswith respect to creepage for all track conditions nonlinearly. In the second step, the results of the simulation in MATLAB and the estimation of the variables mentioned in the previous section are shown and discussed.

Table 1: Polach model parameters under different friction conditions [24]

Model parameter	Wheel–rail conditions			
	Dry	Wet	Low	Very Low
k_A	1	1	1	1
k_S	0.4	0.4	0.4	0.4
D	0.6	0.2	0.2	0.1
B	0.4	0.4	0.4	0.4

The parameter values for the traction system employed in this research can be found in Table 2.

Matrices Q and R are given in the following, which are obtained by trial and error.

Q = diag ([3.88e-7 1.00e-12 1.39e-16 1.42e-16 1.85e-10 1.85e-3]) × 0.099, R = diag ([3.39e-4 3.39e-4]) × 2.

Table 2: Parameters and values used in the simulation

C_v ($\frac{N.m}{rad.s}$)	0.015	J_{eqv} (kg.m^2)	0.07
L_S(H)	0.1004	R_S (Ω)	1.54
L_m(H)	0.0915	R_r (Ω)	1.294
L_r(H)	0.0969	r (m)	0.34
n_i	7.5	f (Hz)	50
n_p	3		

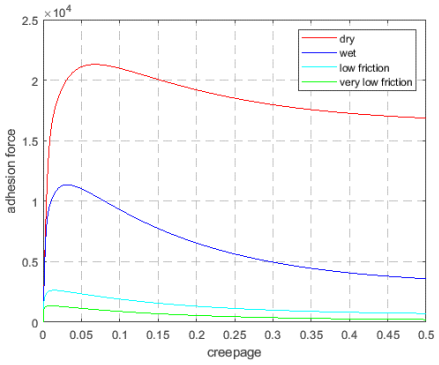


Fig. 3: Adhesion force curves creepage.

Estimated and actual trajectories of stator currents in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ frames ($\hat{\boldsymbol{I}}_{s\alpha}$,$\hat{\boldsymbol{I}}_{s\beta}$) are shown in Fig. 4 (a) and (b), respectively. The trajectories of the current errors in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ frames ($\boldsymbol{e}_{I_{s\alpha}}$,$\boldsymbol{e}_{I_{s\beta}}$) are represented in Fig. 5 (a) and (b), respectively.

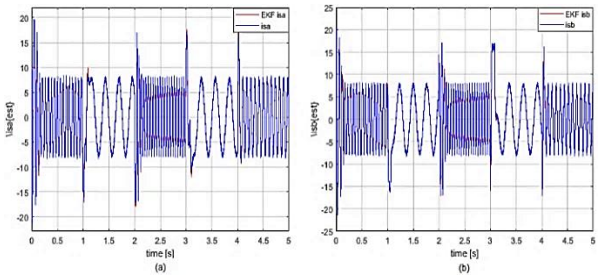


Fig. 4: EKF based estimated and actual motor current trajectories (a) α axis (b) β axis.

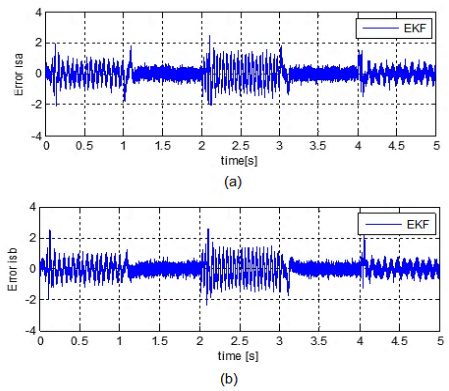


Fig. 5: The trajectories of the motor current error (a) α axis (b) β axis.

Estimated and actual trajectories of rotor fluxes in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ frames ($\hat{\boldsymbol{\psi}}_{r\alpha}$, $\hat{\boldsymbol{\psi}}_{r\beta}$) are represented in Fig. 6 (a) and (b), respectively. The trajectories of the rotor flux errors in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ frames ($\mathbf{e}_{\psi_{s\alpha}}$,$\boldsymbol{e}_{\psi_{s\beta}}$) are represented in Fig. 7 (a) and (b), respectively.

As seen in Fig. 4 (a) and (b) and Fig. 6 (a) and (b), the estimated trajectories of the stator current and rotor flux in α and β frames follow the real trajectories of these four motor variables with minimal error bound.

Estimated and actual trajectories of motor speed (ω_m, $\hat{\omega}_m$) and speed error (e_{ω_m}) are shown in Figs. 8 and 9 respectively. In Fig. 8, fast convergence with a very low bound of error in following the real trajectory by the EKF estimator is clearly evident. The trajectories of the estimated and actual load torque (T_L, $\hat{T}_L$) and load torque error (e_{T_L}) are given in Figs. 10 and 11 respectively.

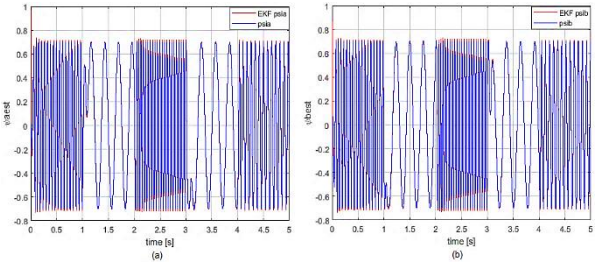


Fig. 6: EKF based estimated and actual rotor flux trajectories (a) α axis (b) β axis.

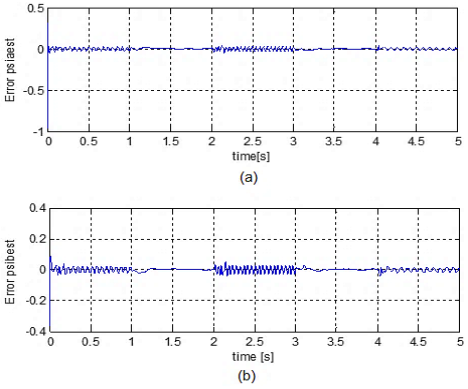


Fig. 7: The trajectories of the rotor fluxes error (a) α axis (b) β axis.

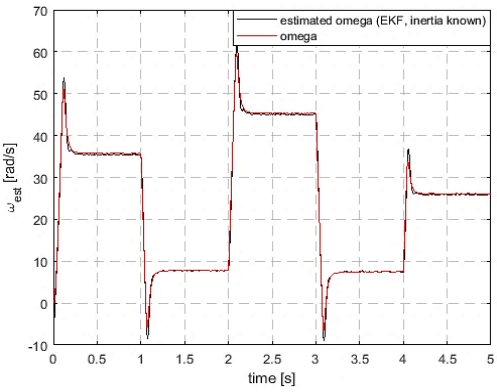


Fig. 8: EKF based estimated and actual trajectories of motor speed.

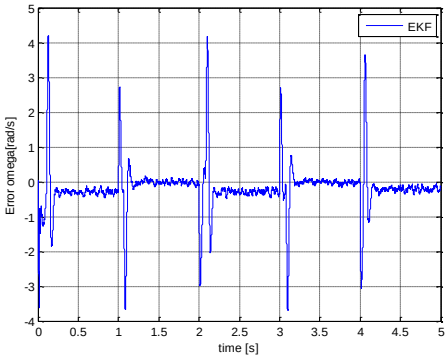


Fig. 9: The trajectory of the motor speed error.

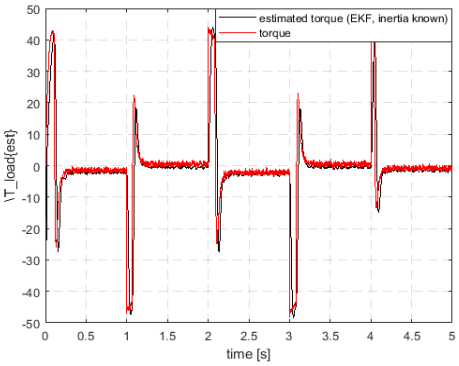


Fig. 10: EKF based estimated and actual trajectories of load torque.

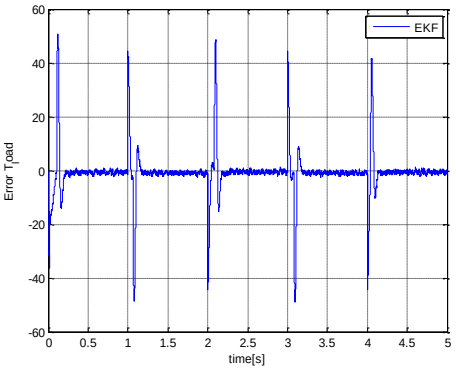


Fig. 11: The trajectory of the load torque error.

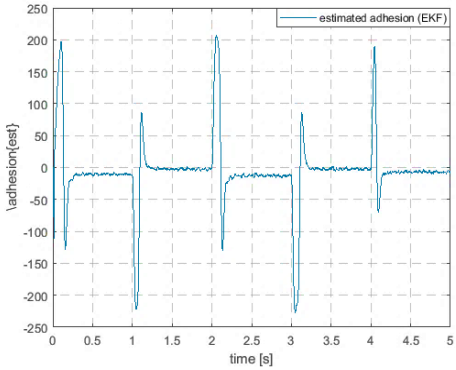


Fig. 12. The trajectory of the estimated adhesion force.

Referring to Fig. 10, it is evident that the error undergoes a narrow variation when there is a sudden change in the torque command. By analyzing Figs. 4 to 12, it becomes apparent that the EKF estimator accurately follows the real state trajectories with great precision and rapid convergence. Equation (18) allows us to obtain the trajectory of the estimated adhesion force, demonstrating a linear correlation between the load torque and the adhesion force. Fig. 12 shows the estimated adhesion force trajectory. By estimating the longitudinal creep force, it becomes feasible to ascertain the degree of adhesion between the wheel and the rail.

B. Low Speed Performance

To further investigate the proposed method at low speeds, the findings of the estimated stator current, rotor flux, motor speed, and load torque compared to the actual conditions are illustrated in Figs. 13-16. Upon analyzing Figs. 15 and 16, it is evident that the proposed method showcased in this study offers prompt response and precise estimation of speed and torque across the entire low-speed range. In Fig. 15, the reference speed is initially set to 6 rad/s, then altered to -6 rad/s at 4s, and finally adjusted back to 6 rad/s at 14s. Fig. 16 indicates minor estimation errors of similar magnitude. The estimated load torque demonstrates the successful operation of the proposed scheme.

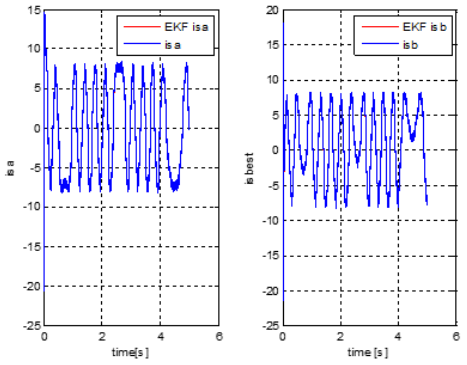


Fig. 13: EKF based estimated and actual motor current trajectories (a) α axis (b) β axis at low speed operation.

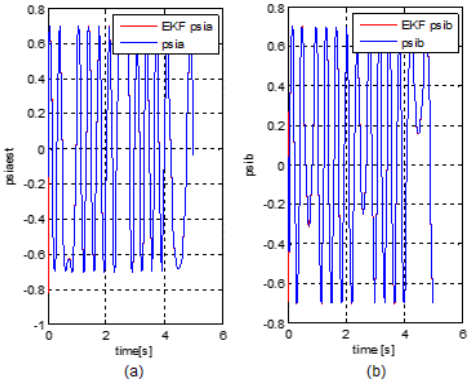


Fig. 14: EKF based estimated and actual rotor flux trajectories (a) α axis (b) β axis at low speed operation.

Fig. 17 displays the trajectory of the adhesion force in low-speed scenarios, which was plotted based on the linear relationship between load torque and adhesion

force. These results confirm the robustness and exceptional tracking capabilities of the estimation approach, even when operating at lower velocities.

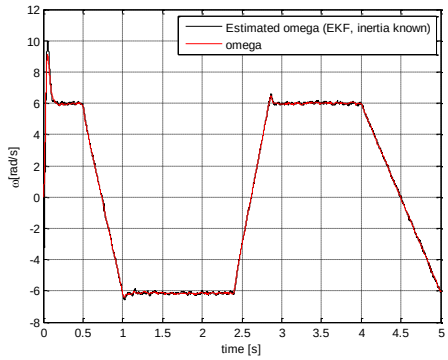


Fig. 15: EKF based estimated and actual trajectories of motor speed at low speed operation.

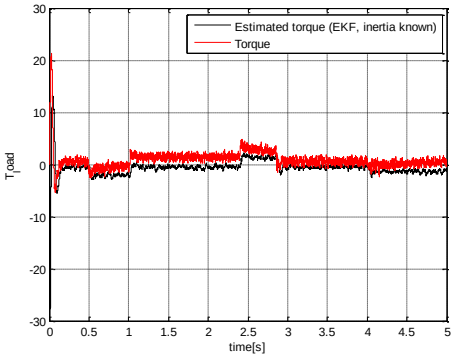


Fig. 16: EKF based estimated and actual trajectories of load torque at low speed operation.

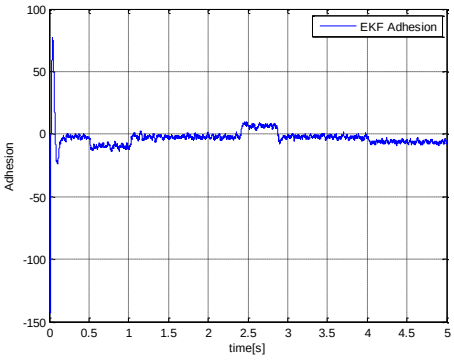


Fig 17: EKF based estimated trajectory of adhesion force at low speed operation.

C. Performance Comparison

As mentioned and considered in subsection 4-1, the estimation of variables with EKF results in good information. For a more detailed investigation and to check the amount of estimation error, and according to subsection 3-2, the performance of the proposed EKF is compared with UKF, validated with Matlab simulation, and comparative analysis is discussed.

In Figs. 18-21 the estimation results of motor current, stator flux, motor speed, and load torque are represented. Both EKF and UKF estimator are used for estimating variables and their outputs compared with the actual situation. As it can be seen, the obtained results present a high degree of convergence, acceptable accuracy, and good estimation of variables in estimating with the EKF estimator. The trajectory of the estimated adhesion in two estimation modes i.e. EKF and UKF is given in Fig. 22. Based on the analysis of the estimation results, it can be deduced that the UKF algorithm, known for its effectiveness in highly-nonlinear systems as per previous research, does not exhibit any advantage over the EKF algorithm when it comes to estimating induction motor parameters and statistics.

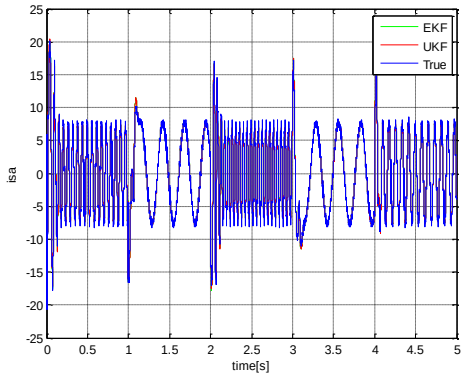


Fig. 18: EKF and UKF based estimated and actual motor current trajectories.

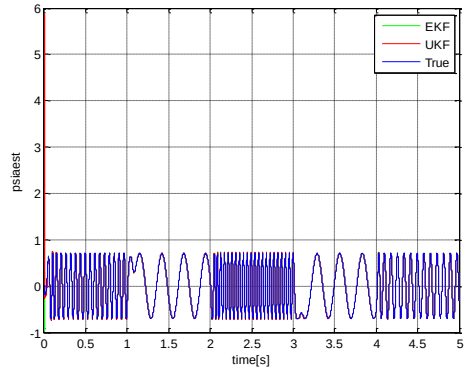


Fig. 19: EKF and UKF based estimated and actual rotor flux trajectories.

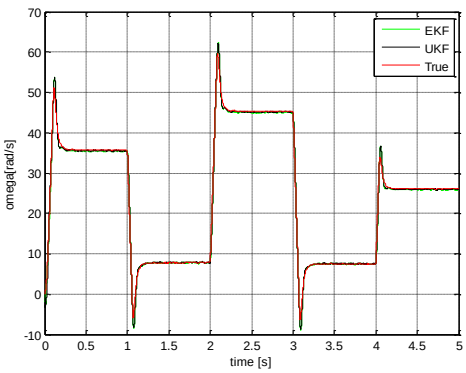


Fig. 20: EKF and UKF based estimated and actual motor speed trajectories.

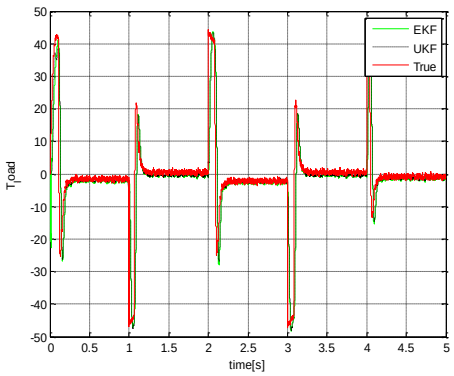


Fig. 21: EKF and UKF based estimated and actual load torque trajectories.

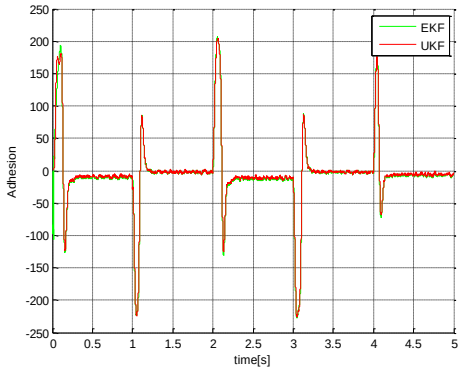


Fig. 22: EKF and UKF based estimated adhesion trajectories.

In the following, the trajectories of the estimation error for all mentioned state variables with the UKF and EKF estimation modes are shown in Figs. 23-26. By comparing the obtained results, it is obvious that EKF can estimate variables with high accuracy and partial error in the presence of UKF.

Fig. 26 shows the load torque error in estimating with both EKF and UKF estimators. According to the (18), there is a linear relation between load torque and adhesion force.

Hence, it can be deduced that the EKF exhibits swift responsiveness and provides estimations with minimal margin of error, given the fluctuating adhesion circumstances between the wheel and rail surfaces.

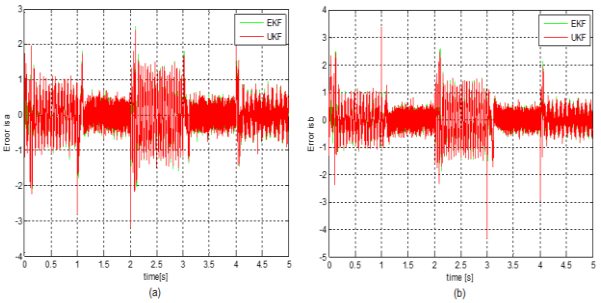


Fig. 23: The trajectories of the motor current error (a) α axis (b) β axis.

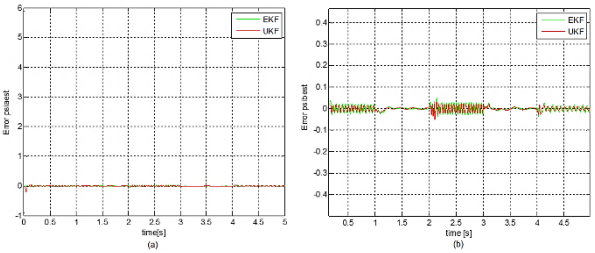


Fig. 24: The trajectories of the rotor flux error (a) α axis (b) β axis.

To further evaluate the estimation accuracy of the approach, the root mean square error (RMSE) of state variables is shown in Figs. 27-30. By analyzing the Figs, it can be concluded that the RMSE of EKF is smaller than the RMSE of UKF, as a result, the speed estimated by EKF is closer to its actual values.

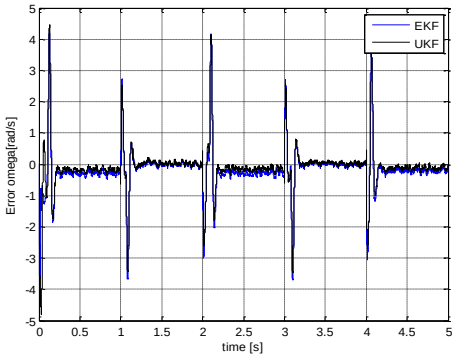


Fig. 25: The trajectory of the motor speed error.

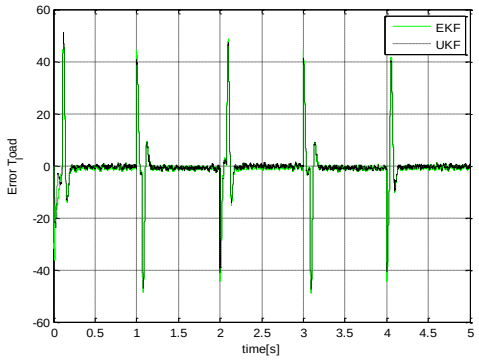


Fig. 26: The trajectory of the load torque error.

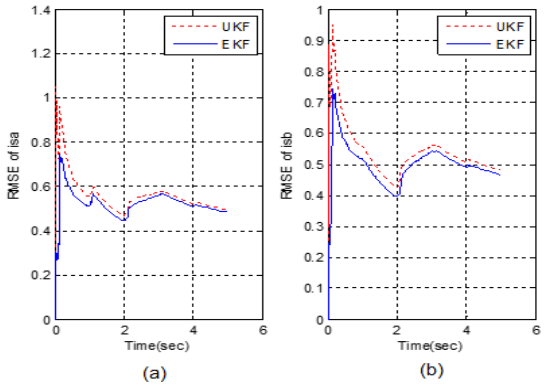


Fig. 27: The RMSE of motor current over time (a) in α axis (b) in β axis.

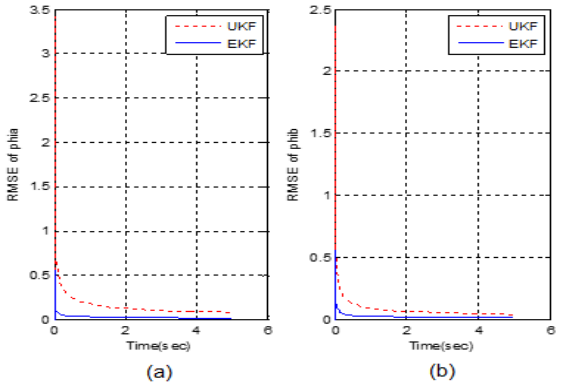


Fig. 28: The RMSE of stator flux over time (a) α axis (b) β axis.

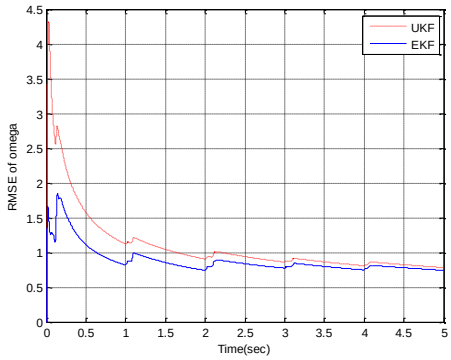


Fig. 29: The RMSE of motor speed over time.

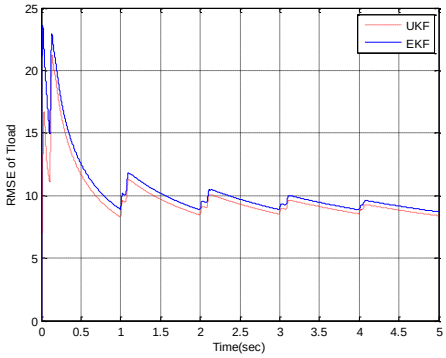


Fig. 30: The RMSE of load torque over time.

Table 3. Running times of EKF and UKF algorithms

Estimator	Execution Time (sec)
EKF	2.6
UKE	9.2

It can be seen in Fig. 30 that the RMSE of load torque with EKF is equal to that of UKF. Therefore, it is the same for adhesion force due to the linear relationship between the load torque and adhesion force. In order to show the difference clearly and better understand the two algorithm performances, the computational cost of methodes is given in Table 3. The results show that the running time of EKF is lower than UKF, which confirm the superiority of EKF relative to UKF in estimating adhesion force and induction motor parameters.

Conclusion

In this research, an EKF-based condition monitoring is proposed to estimate adhesion force. To assess the efficiency of the estimator, an evaluation was conducted by comparing the estimated motor parameters such as load torque, speed, rotor flux, and stator current in three modes i.e. actual, EKF-based, and UKF-based. Meanwhile, the linear relationship between motor torque and adhesion force was used to determine the adhesion level between the wheel and the rail. The results indicated that the EKF estimator demonstrates prompt responsiveness and accurately estimates the variables, despite the varying adhesion conditions of the wheel-rail contact. The estimator consistently maintains a minimal margin of error. The remarkable thing is that EKF shows its superiority in state and parameter estimation of induction motor and adhesion force since UKF is not able to exhibit its effectiveness for this type of application. Therefore, utilizing such estimator can help to achieve maximum traction, reduce the creepage, and improve the performance of the re-adhesion controller. One major issue with the EKF and UKF lies in the significant impact of the covariance matrices Q and R on the estimation outcomes. If these matrices are chosen badly, the estimation result will be divergent or large estimate errors will be inevitable.

Tuning of Q and R is necessary to yield the best estimations. Changing these two matrices affects both the steady-state and transient duration operation of the estimator. When the value of Q increases, it indicates the presence of significant disturbances or uncertainties in the machine model. Consequently, the Kalman gain is augmented, resulting in a faster performance of the estimator during the transitional phase. On the other hand, increasing R means that noise measurements are strong and the noise will be weighted less by estimator, which leads to Kalman gain decrease and gives us a slower transient performance. To enhance the performance, accuracy, and stability of the estimator, it is imperative to employ various tuning algorithms. In our upcoming endeavors, we will focus on implementing these algorithms and give particular attention to developing a real-time system.

Author Contributions

M. Moradi collected the data, carried out the analysis and wrote paper, R.Havangi wrote the paper ,interpreted the results and supervised the research.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

a and b: Semi-axis length of the contact patch

B and D: Reduction factors

C_{11}: Kalker coefficient

C_v: Viscous friction

F_a: Adhesion force

F_N: Normal force between the wheel and rail

G: Shear module

$I_{s\alpha}$ and $I_{s\beta}$: $\alpha-\beta$ axis stator currents

c_1 and c_2: Self-recognition and social component coefficients

J_{eqv}: Equivalent moment of inertia

J_g: Gearbox moment of inertia

J_x: Wheelset axle moment of inertia

$J_{wR}\ and\ J_{wL}$: Right and left wheel moment of inertia

k_A and k_S: Reduction factors in the adhesion and slip area

L_m: Mutual inductance

L_r and L_s: Rotor and stator self-inductance

n_i: Gear reduction ratio

n_p : Number of the pole pairs

N: Number of unknown variables or number of samples

P_i: Previous best position of each particle

Q and R: Process and measurement noise covariance matrixes

r: Wheel radius

R_r and R_s: Rotor and stator resistance

n_i: Gear reduction ratio

T_m: Motor torque

T_L: Load torque

V: Longitudinal velocity

$v(t)$ and $w(t)$: Measurement and process noise

V_i: Particle velocity

w: Inertia weight factor

X_i: i^{th} particle position

$\psi_{r\alpha}$ and $\psi_{r\beta}$: $\alpha-\beta$ axis rotor flux

μ_f: Traction coefficient

σ: Leakage coefficient

ϵ: Gradient of tangential stress

ξ: Creepage between the wheel and rail

ω_m: Motor angular velocity

ω_w: Wheel angular velocity

References

[1] C. Schwarz, A. Keck, “Simultaneous estimation of wheel-rail adhesion and brake friction behaviour,” IFAC, 53: 8470-8475, 2020.

[2] R. Lewis, G. Trummer, K. Six, J. Stow, H. Alturbeh, P. Shackleton, B. Bryce, L. Buckley Johnstone, “Leaves on the line: Characterizing leaf based low adhesion on railway rails,” Tribol. Int., 185: 108529, 2023.

[3] H. Chen, T. Furuya, S. Fukagai, S. Saga, J. Ikoma, K. Kimura, J. Suzumura, “Wheel slip/slide and low adhesion caused by fallen leaves,” Wear, 446-447: 203187, 2020.

[4] M. Watson, B. White, J. Lanigan, T. Slatter, R. Lewis, “The co-mposition and friction reducing properties of leaf layers,” Proc. Math. Phys. Eng. Sci., 476 (2239), 2020.

[5] J. Zhou, M. Wu, C. Tian, et al., “Experimental investigation on wheel–rail adhesion characteristics under water and large sliding conditions,” Ind. Lubric. Tribol., 73(2): 366-372, 2021.

[6] Y. Lyu, E. Bergseth, U. Olofsson “Open system tribology and influence of weather condition,” Sci. Rep. , 6: 32455, 2016.

[7] M. Harmon, R. Lewis “Review of top of rail friction modifier tribology,” Tribol. Mater. Surface Interfac., 10(3): 150-162, 2016.

[8] M. Shen., Y. Qin, D. Ji, et al. “Role of ambient temperature in the adhesion and damage characteristics of wheel/rail interface during rolling-sliding contact,” Wear, 506-507: 204458, 2022.

[9] H. Chen, H. Tanimoto, “Experimental observation of temp-erature and surface roughness effects on wheel/rail adhesion in wet conditions,” Int. J. Rail Transport., 6(2): 101-112, 2018.

[10] U. Olofsson, Y. Lyu, “Open system tribology in the wheel–rail contact—a literature review,” Appl. Mech. Rev. 69(6): 60803, 2017.

[11] I. Yasuoka, T. Henmi, Y. Nakazawa, I. Aoyama, "Improvement of re-adhesion for commuter trains with vector control traction inverter," in Proc. Power Conversion Conference - PCC '97, 1: 51-56, 1997.

[12] Y. Matsumoto, N. Eguchi, A. Kawamura, “Novel re-adhesion control for train traction system of the "Shinkansen" with the estimation of wheel-to-rail adhesive force,” in Proc. IECON'01. 27th Annual Conference of the IEEE Industrial Electronics Society (Cat. No.37243), 2: 1207-1212, 2001.

[13] K. Zhao, P. Li, Ch. Zhang, J. He, Y. Li, T. Yin, “Online accurate estimation of the wheel-rail adhesion coefficient and optimal adhesion antiskid control of heavy-haul electric locomotives based on asymmetric barrier lyapunov function,” J. Sensors, 2018: Article ID 2740679, 2018.

[14] R. Bibi, B. S. Chowdry, R. A. Shah, “PSO based localization of multiple mobile robots employing LEGO EV3,” in Proc. 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET): 1-5, 2018.

[15] S. Shrestha, Q. Wu, M. Spiryagin, “Review of adhesion es-timation approaches for rail vehicles,” Int. J. Rail Transport., 7(2): 79-102, 2019.

[16] X. Fang, S. Lin, Z. Yang , F. Lin, H. Sun, L. Hu “Adhesion control st-rategy based on the wheel-rail adhesion state observation for high-speed trains,” Electronics, 7(5): 70, 2018.

[17] B. Liu, T. X. Mei, S. Bruni, "Design and optimisation of wheel–rail profiles for adhesion improvement," Veh. Syst. Dyn., 54(3): 429-444, 2016.

[18] Y. Chen, H. Dong, J. Lu, X. Sun, L. Guo, "A super-twisting-like algorithm and Its application to train operation control with optimal utilization of adhesion force," IEEE Trans. Intell. Transpor. Syst., 17(11): 3035-3044, 2016.

[19] M. Yamashita, T. Soeda, "Anti-slip re-adhesion control method for increasing the tractive force of locomotives through the early detection of wheel slip convergence," in Proc. 17th European Conference on Power Electronics and Applications: 1-10, 2015.

[20] R. Rizzo, D. Iannuzzi, "Indirect friction force identification for application in traction electric drives," Math. Comput. Simul., 60(3-5): 379-387, 2002.

[21] I. Hussain, T. X. Mei, R. T. Ritchings, "Estimation of wheel– rail contact conditions and adhesion using the multiple model approach," Veh. Syst. Dyn., 51(1): 32-53, 2013.

[22] G. Charles, R. Goodall, R. Dixon, "Model-based condition monitoring at the wheel-rail interface," Veh. Syst. Dyn., 46(1): 415-430, 2008.

[23] F. Orderud, "Comparison of kalman filter estimation approa-ches for state space models with nonlinear measurements," in Proc. Scand. Conf. Simul., 194: 157-162, 2005.

[24] Y. Zhao, B. Liang, "Re-adhesion control for a railway single wheelset test rig based on the behaviour of the traction motor," Veh. Syst. Dyn., 51(8): 1173-1185, 2013.

[25] S. Wang, J. Xiao, J. Huang, H. Sheng, "Locomotive wheel slip detection based on multi-rate state identification of motor load torque,"J. Franklin Inst., , 353(2): 521-540, 2016.

[26] P. D. Hubbard, C. Ward, R. Dixon, R. Goodall, "Verification of model based adhesion estimation in the wheel-rail interface," Chem. Eng. Trans., 33: 757-762, 2013.

[27] A. Onat, P. Voltr, M. Lata, "An unscented Kalman filter-based rolling radius estimation methodology for railway vehicles with traction," Proc. Inst. Mech. Eng., Part F: J. Rail Rapid Transit, 232(6): 1686-1702, 2018.

[28] S. Jafarzadeh, C. Lascu, M. Fadali, "Square root unscented Kalman Filters for state estimation of induction motor drives," IEEE Trans. Ind. Appl., 49(1): 92-99, 2013.

[29] S. Jafarzadeh, C. Lascu, M. Fadali, "State estimation of induction motor drives using the unscented Kalman filter," IEEE Trans. Ind. Electron., 59(11): 4207-4216, 2012.

[30] B. Akin, U. Orguner, A. Ersak, M. Ehsani, "Simple derivative free nonlinear state observer for sensorless AC drives," IEEE/ASME Trans. Mech., 11(5): 634-643, 2006.

[31] M. Barut, S. Bogosyan, M. Gokasan, "Speed-sensorless estimation for induction motors using extended Kalman filters," IEEE Trans. Ind. Electron., 54(1): 272-280, 2007.

[32] M. Spiryagin, P. Wolfs, C. Cole, V. Spiryagin, "Design and simulation of heavy haul locomotives and trains," CRC Press, Boca Raton, FL, USA, 2016.

[33] O. Polach, "A fast wheel-rail forces calculation computer," Veh. Syst. Dyn. Suppl, 33(1): 728-739, 1999.

[34] J. Kalker, "On The rolling contact of two elastic bodies in the presence of dry friction," Wear, 11: 303, 1967.

Biographies

Maryam Moradi received her B.S. degree in Control Engineering from Gonabad Azad University, Gonabad, Iran in 2008, and her M.S. degree in Telecommunications Engineering from the Faculty of Engineering at the University of Sistan and Baluchestan, Zahedan, Iran in 2015. Since 2018, she has been a Ph.D. student in Control Engineering at the Faculty of Engineering, the University of Birjand, Birjand, Iran. In 2019, she joined the University of Applied Sciences & Technology as a teacher.

- Email: m_moradi@birjand.ac.ir
- ORCID: 0009-0007-0635-6867
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Ramazan Havangi received his M.S. and Ph.D. degrees from the K.N. Toosi University of Technology, Tehran, Iran in 2003 and 2012, respectively. He is currently an Associate Professor of control systems with the Department of Electrical and Computer Engineering at the University of Birjand, Birjand, Iran. His main research interests are inertial navigation, integrated navigation, estimation and filtering, evolutionary filtering, simultaneous localization and mapping, fuzzy, neural network, and soft computing.

- Email: : Havangi@Birjand.ac.ir
- ORCID: 0000-0001-5711-3127
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://cv.birjand.ac.ir/havangi/fa

How to cite this paper:
M. Moradi, R. Havangi, "Estimation of wheel-rail adhesion force using traction system behavior," J. Electr. Comput. Eng. Innovations, 12(1): 271-282, 2024.

DOI: 10.22061/jecei.2023.9935.664

URL: https://jecei.sru.ac.ir/article_2014.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

An Efficient Region-of-Interest (ROI) based Scalable Framework for Free Viewpoint Video Application

H. Roodaki *

Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran.

Article Info

Article History:
Received 24 August 2023
Reviewed 02 October 2023
Revised 06 November 2023
Accepted 12 December 2023

Keywords:
Tile-based scalability
Region of interest
λ-domain rate control algorithm
MV-HEVC
Parallel processing

*Corresponding Author's Email Address: *hroodaki@kntu.ac.ir*

Abstract

Background and Objectives: From the multiview recorded video, free viewpoint video provides flexible viewpoint navigation. Thus, a lot of views need to be sent to the receivers in an encoded format. The scalable nature of the coded bitstream is one method of lowering the volume of data. However, adhering to the limitations of the free viewpoint application heavily relies on the kind of scalable modality chosen. The perceptual quality of the received sequences and the efficiency of the compression technique are significantly impacted by the scalable modality that was chosen.

Methods: In order to address the primary issues with free-viewpoint video, such as high bandwidth requirements and computational complexity, this paper suggests a scalable framework. The two components of the suggested framework are as follows: 1) introducing appropriate scalable modality and data assignment to the base and enhancement layers; and 2) bit budget allocation to the base and enhancement layers using a rate control algorithm. In our novel scalable modality, termed Tile-based scalability, the idea of Region of Interest (ROI) is employed, and the region of interest is extracted using the tile coding concept first presented in the MV-HEVC.

Results: When compared to the state-of-the-art techniques, our approach's computational complexity can be reduced by an average of 44% thanks to the concept of tile-coding with parallel processing capabilities. Furthermore, in comparison to standard MV-HEVC, our suggested rate control achieves an average 17.7 reduction in bandwidth and 1.2 improvement in video quality in the Bjøntegaard-Bitrate and Bjøntegaard-PSNR scales.

Conclusion: Using new tile-based scalability, a novel scalable framework for free-viewpoint video applications is proposed. It assigns appropriate regions to the base and enhancement layers based on the unique features of free viewpoint scalability. Next, a rate control strategy is put forth to allocate a suitable bitrate to both the base and enhancement layers. According to experimental results, the suggested method can achieve a good coding efficiency with significantly less computational complexity than state-of-the-art techniques that used the λ-domain rate control method.

Introduction

Free viewpoint video is a system allowing the users to observe the scene from various view point and freely change the views. It can be used in multiple applications such as immersive teleconference, 3DTV, and sporting

events, enabling the viewers to move freely around the scene. But, to provide this capability, some main challenges must be overcome.

- **High Bandwidth:** To provide high quality and rich interactive experience, the resolution of the free viewpoint video should be 4K or even higher. However, streaming numerous views at such resolution requires extraordinary bandwidth, which is not reachable in low bit rate communication channels such as the wireless mobile networks. Even with the compression offered by Multiview Video Coding (MVC), which uses Intra and Inter-view prediction to extract statistical dependencies between views, free viewpoint video is currently beyond the capabilities of most wireless networks. Intra-view prediction uses temporally adjacent frames, and inter-view prediction uses corresponding frames in adjacent views as reference views in the prediction process.
- **Computational Complexity:** As mentioned before, an essential characteristic of free viewpoint video is increasing the control over view angle and direction, according to the perceptual preferences of users. A secondary effect of increasing the viewing controllability is higher computational complexity. Since, there are numerous numbers of views that should be encoded and send to the viewers, while the human eye can only focus on a particular area of the scene and small numbers of view at any time. Hence, the main challenge in this application is controlling the high computational complexity of encoder side [1].

Scalable Multiview Video Coding (SMVC) is one of the main techniques to address the high bandwidth requirement and reduce computational complexity by scaling down the video. Usually, a scalable bitstream consists of a “Base” layer that carries the minimum amount of video data necessary for all receivers. Then, one or more “Enhancement” layers can be built on top of the base layer to improve the overall video perceptual quality.

In this paper, we propose an efficient scalable framework for free viewpoint video applications to overcome the mentioned challenges. Our proposed framework has two distinct parts:

1) Selecting proper scalable modality and data assignment to base and enhancement layers: Selecting the proper scalable modality can help satisfying the main requirements of free-viewpoint application such as the required bandwidth. In particular, the selected scalable modality can affect the perceived quality and the compression efficiency.

To clarify the meaning of scalable modality, suppose that, a mobile client has an available bandwidth that is less than the bitrate of the original free viewpoint video. A video adaption should be performed to extract base and enhancement layers to match the bitrate of the adapted video to the bandwidth of the mobile client. This type of video adaptation is named scalable modality. It is more efficient that the scalable modality is extracted according to the specific characteristics of the application at hand, such as free viewpoint application. After selecting a proper scalable modality, the available data should be assigned to various layers according to this particular scalable modality. In the free viewpoint application, usually, the receivers are more interested in specific regions of the scene, and the perceptual quality of those regions is more important to them. This concept is typically referred to as Region of Interest (ROI). Our proposed framework in this paper proposes to use the concept of ROI to define the appropriate scalable modality for free viewpoint application. The user's favorite regions, ROIs, are located in the base layer and the other areas are located in enhancement layers.

But, extracting the ROI is not so trivial. Several methods are presented in the literature to extract ROI based on some specific features such as texture perceptual map and the motion perceptual map [2], visual attention model, and gaze tracking data [3] and so on.

In this paper, we suggest using the idea of tile coding in the MV-HEVC standard to extract the ROI and data assignment to the base and enhancement layers.

In MV-HEVC standard, the pictures can be divided into independently decodable rectangular regions with approximately equal numbers of CTUs that are entitled as tiles. The main goal of partitioning the frames into tiles is to increase the capability for parallel processing and provide error resilience [4].

Using this concept, we can code the tiles corresponding to ROI as the base layer and the remaining ones as the enhancement layers. The tiles can be coded and decoded in parallel to improve the computational complexity of the encoder and decodes side.

2) Rate control algorithm for bit budget allocation to base and enhancement layers: For the second part of the proposed framework, we propose a rate control algorithm for scalable coding in free viewpoint application. We have used the λ-domain rate-distortion model [5] in the HEVC video coding standard as our reference rate-distortion model. Then we will use specific features of the Tile-based scalable modality to extract coding parameters for free viewpoint video application efficiently.

The main innovations of the paper are as follows:

1. Introduce a proper scalable modality for free-viewpoint video application based on Region of Interest, named Tile-Based scalability.
   - In tile-based scalability, the concept of tile coding in the MV-HEVC video coding standard is used to extract the interested and non-interested regions

and allocate them to the base and enhancement layers, respectively.
- In addition, using the concept of tile coding in the MV-HEVC video coding standard, we can benefit from parallel processing to reduce the complexity of the encoder and improve efficiency.

2. Propose a rate control algorithm for rate assignment to base and enhancement layers for Tile-Based scalability in free-viewpoint video application.
   - The concept of inter-view disparity is used to find the appropriate relationship between the quantization parameters of various tiles

Related Work

In this section, the scalable coding for multiview video and free viewpoint application and the rate control algorithms for scalable coding are reviewed.

A. Scalable Coding in Multiview Video and Free Viewpoint Application

Several scalable modalities have been mentioned in the literature for single view and multiview video. For instance, temporal, spatial, and quality scalability and various combinations of them [6] and Region-Of-Interest (ROI) and object-based scalability [7] are introduced for single view video coding. Besides, view scalability [8] and free viewpoint scalability [9] are presented for multiview video. In [10] an efficient scalable multiview video coding method is proposed in which high-quality depth maps are coded as a piece of scene information. Then, the view-dependent depth map is generated from this information at the decoder side. So the free viewpoint scalability and coarse granular SNR scalability are achieved using these synthesized depth maps. In [11], a scalable approach is presented for immersive video streaming to support different receivers. This method limits the number of views in the base layer and uses view scalability and free viewpoint scalability in the enhancement layers to synthesize more views at the receiver side to improve the quality of free viewpoint views for the user. In [12] an encoding configuration for scalable multiview video coding is proposed that realizes higher compression efficiency and provides view switching for the users. In the proposed approach, the base layer uses inter-view prediction and produces a video sequence with the acceptable quality. Then, the enhancement layers use the corresponding base layer without interview prediction.

As we mentioned before, complying the constraint of free viewpoint application is strongly dependent on the type of selected scalable modality that has a significant effect on the effectiveness of the compression method and perceptual quality of received sequences. The above mentioned scalable modalities are not fitted precisely to the particular characteristics of free viewpoint application. Hence, in this paper, we propose a new scalable modality according to the specific features of this application.

B. Rate Control Algorithms in Multiview Scalable Video Coding

In this section, the most recent rate control algorithms for multiview scalable video coding are presented. In [13] a ρ-domain rate control algorithm for multiview high efficiency video coding is proposed. First, the prediction structure of MV-HEVC is optimized. Then, the ρ domain rate control model based on multi-objective optimization is used. In this study, the image similarity is considered to analyze the correlation between viewpoints. Then, this correlation and the frame complexity are used for the rate allocation process. The method proposed in [14] uses the analysis of the characteristics of multi-view video coding and the requirements of its bit rate control to improve the traditional quadratic rate-distortion model. In [15], a bit allocation and rate control approach for multiview video coding is proposed that uses the frame complexity and human visual characteristics. These characteristics are used to improve the quadratic rate-distortion (R-D) model. Then, the proposed algorithm reasonably allocates bit-rate among views based on frame complexity and human visual characteristics. The bit allocation process among various views is done by solving a multi-objective optimization problem. In [16], a rate control algorithm for MV-HEVC based on scene detection is proposed. A ρ-domain rate control model based on multi-objective optimization that uses the image similarity is used. This image similarity is suggested to make a reasonable bit allocation among viewpoints. So, by switching the video scene, the image similarity is recalculated. In this paper, the frame layer rate control considers the layer B-frame and other factors in allocating the code rate. Then, the basic unit layer rate is done according to the content complexity of the CTU. In [17], a three levels quadratic rate-distortion model for multiview video coding is proposed. A λ-domain rate control algorithm for the scalable extension of HEVC video codec is presented in [18] that includes temporal, spatial, and quality scalability. The proposed algorithm introduces an initial target bits and encoding parameters determination algorithm for the first frame of each layer. Then, considering the inter frames, a bit allocation method is suggested using intra and inter layer dependencies. In [19] a rate control method is proposed that takes the human visual system into account to allocate the bit budget to various vision perceptual regions. The proposed model in [20] takes the QP values of B frames into account since some views in the multiview video are consisted of B frames only. In [21] a rate-distortion model is introduced that considers the dependency between frames made by motion compensation and depth image-based rendering. Finally,

[22] proposes a method for key frames that encouraged R-λ model using the depth map characteristics.

A novel method for rate assignment of free-viewpoint video is presented in [23] that uses the distance between view directions to allocate the appropriate rate for each view and provide a broader field of view.

In [24] rate control algorithm for MV-HEVC based on scene detection is presented. The motivation is that the rate control algorithm for multiview in (MV-HEVC) does not have the capability of bit allocation efficiently at the CTU level. So, the video quality varies greatly for sequences with sudden scene changes or large motions.

As we can see, most of the above mentioned methods are suggested for multiview video and cannot be extended to scalable multiview video easily. Also, besides considering the main features of scalable modality in the rate allocation process may lead to much better performance in the compression process. To the best of our knowledge, none of the presented rate allocation methods addresses this issue.

Proposed Method

In this section, we introduce our proposed scalable framework for free viewpoint application. Our proposed framework has two distinct steps, 1) introducing a new scalable modality for free viewpoint application and 2) presenting a rate allocation method to assign reasonable rate to base and enhancement layers.

A. Introduce A New Scalable Modality for Free Viewpoint Application

Free viewpoint video often uses in situations that want to give the viewers the higher coverage of Field of View. Hence, the number of cameras and the corresponding setup should provide the most coverage from the scene. But, it is usually more important to the viewers to see the most interesting parts of the scene than the events around the boundaries. For instance, a football game often relies on the movement of participants within a specific playing area. Where the interesting events will take place, are more important to the viewers. Missing these areas will impose a harmful effect on the user's quality of experience [25].

Our proposed scalable modality for free viewpoint application suggests using this criterion to assign data to the base and enhancement layers. The most interesting parts of the scene that are more important to the users are located in the base and the other parts are located in one or more enhancement layers. But, the main challenge is how to extract the interesting parts of the scene.

The proposed ROI selection methods are either feature-based or object-based. Feature-based methods find pixels that share significant optical features with the target and aggregate them to form ROIs [26], [27]. These methods can capture most of the target pixels based on the optical feature similarity. However, not all target pixels have strong optical features, so the detected ROI usually fails to encompass the entire target. In addition, feature-based methods cannot distinguish between targets, which can cause confusion in subsequent stages of processing.

Object-based methods, on the other hand, detect ROIs at a higher level than the pixel-by-pixel approach of feature-based systems using information such as target shape and structure [28], [29]. Typical approaches include template matching and matched filters. Although these methods can assign a single ROI to one target, they are limited because they require many calculations, have difficulty detecting multiple target types, and are not reliable when applied to low-quality images.

According to the above discussion, finding the ROI using feature-based and object-based methods is computationally complex. In these approaches, various parts of the scene should be traversed pixel-by-pixel or some other information such as shape and structures should be considered.

The concept of tile in MV-HEVC offers an alternative partitioning of a picture into rectangular parts that are encoded and decoded independently of another tile [26]-[30]. The main goal of using tiles coding is to enable the use of parallel processing for the encoding and decoding process [4]. In addition, tile coding facilitates video coding based on ROI. For instance, the tile containing the ROI can be extracted quickly and processed more efficiently [30].

We have used this concept in our proposed framework to introduce a new scalable modality for free viewpoint video, named "Tile-based Scalability". In the proposed scalable modality, in various views, the tiles corresponding to the ROI will be coded independently as the base layer and the other tiles of views corresponding to non-interesting parts are encoded as one or more enhancement layers. This way, selecting the tiles related to the region of interest is simple, since it is not required to detect specific objects or areas. Just finding the approximate coordinates of the desired area is enough. So, this method is much less complex than the feature-based and object-based ROI detection methods. As discussed, this new scalable modality is precisely matched to the free viewpoint application and its requirements.

As discussed before, in free view-point application, numerous numbers of views should be encoded and send to the viewers that increase the computational complexity of the encoder side. In addition, MV-HEVC has developed from previous standards adding some advanced features to increase compression ratios. This higher coding efficiency is obtained at the expense of substantial growth in computational complexity [31]. Besides, the motion compensation and loop filtering functions are the most complex functions at the decoder

side [32].

The suggested method to overcome the high complexity is supporting parallelism at the encoders and the decoders using tile coding [33]. Motion vector prediction, intra prediction, entropy coding, and reconstruction dependencies are not allowed across a tile boundary [33].

So the tiles can be coded end decoded independently from each other and the encoder or the decoder can process a tile in parallel with the other ones.

In tile-based scalability, we generalize this concept to scalable video coding and the tiles of each layer should be encoded and decoded independently and without using the tiles of the other layers. For instance, suppose that our free viewpoint video has three views, each of them has three tiles as shown in Fig. 2.

Assume that, considering the region of interest in this video, we decided to allocate three middle tiles to the base and the other six tiles to enhancement layers1 and 2, respectively.

Hence, the tiles of the base layer, Tile #12, Tile #22, and Tile #32 should be coded from each other using inter-view coding. Tile #11, Tile #21, and Tile #31 form enhancement layer 1 also can use inter-view prediction for much efficient compression. The tiles of different layers, for instance, the Tile #12 from the base layer and Tile #11 of the enhancement layer 1, cannot use for inter-view prediction. Coding the tiles of each layer independently from the other layer can lead to parallel processing of the tiles of each layer that can improve the computational complexity of the encoder and decoder side.

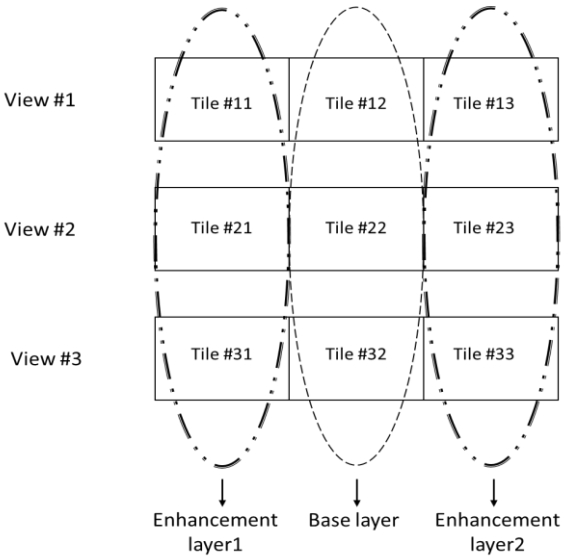


Fig. 2: Tiles allocation to base and enhancement layers in Tile-based scalability in free viewpoint video.

B. Rate Control Algorithms for Tile-Based Scalability in Free Viewpoint Application

Efficient rate allocation to the tile-based scalable video should consist of two steps as shown in Fig. 2. First, assigning the required bitrate to base and enhancement layers according to its importance and then, giving efficient rate to the corresponding tiles of each layer. Our proposed approach uses the main features of tile-based scalability to allocate the proper bitrate to various layers and the corresponding tiles efficiently, as explain in the following sub-sections.

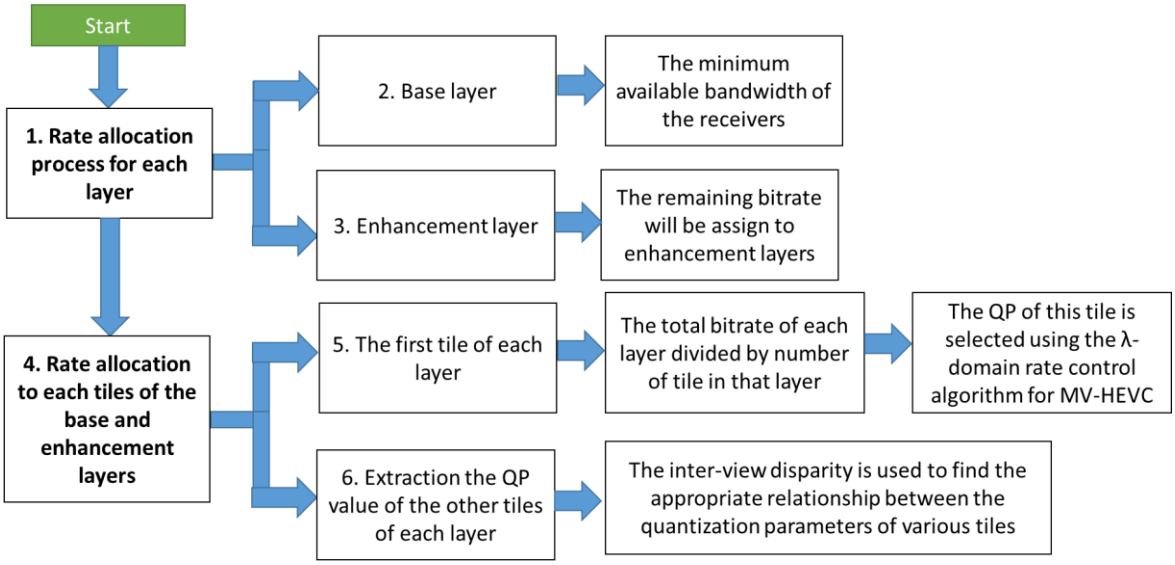


Fig. 1: The block diagram of the proposed method for rate control algorithms.

I) Rate Allocation for Each Layer

For the first step, rate assignment is performed considering the main characteristic of tile-based scalability and free viewpoint application.

In tile-based scalability, the base layer contains the minimum number of required tiles according to the minimum bandwidth of all receivers. Then each specific receiver can request one or more enhancement layers to increase the number of received tiles and to cover the required viewing angle according to its additional bandwidth. Hence, the total bitrate of the base layer should be selected according to the minimum available bandwidth of the receivers. Since, all the receivers should be able to receive this layer. For instance, when a scene is being captured by multiple cameras and should be sent to a mobile phone with limited bandwidth and also a portable tablet with more resources, then, the total bitrate of the base layer should be selected according to the available bandwidth of the mobile phone. Then, the remaining bitrate will be assigned to enhancement layers.

II) Rate Allocation to Each Tiles of the Base and Enhancement Layers

In order to find the bitrate of tiles in each layer, we will use the following steps.

First, an initial bitrate is chosen for the first tile of each layer. This initial bitrate can be considered as the total bitrate of each layer divided by the number of tiles in that layer. Then, the quantization parameter (QP) of this tile is selected using the λ-domain rate control algorithm for MV-HEVC [34].

In λ-domain, the λ parameter is defined as the slope of the R-D curve, which can be expressed as [5]:

$$\lambda = -\frac{\partial D}{\partial R} \triangleq \alpha R^{\beta} \quad (1)$$

Hence,

$$R = \left(\frac{\lambda}{\alpha}\right)^{\frac{1}{\beta}} \quad (2)$$

where α and β are parameters related to the video content and using some pre-encoded video, these parameters values can be extracted via the fitted R –λ curve.

According to (2), the rate-distortion analysis can be carried out in the λ domain and the λ can be determined according to the target bitrate. We have used the initial bitrate for the first tile to extract the λ parameter. Then, the other coding parameters such as QP can be determined using the following equation as suggested in [5].

$$Qp = c1 \times \ln(\lambda) + c2 \quad (3)$$

where c1 and c2 are set to 4.2005 and 13.7122, respectively [34]. The extracted QP value from (3) should be rounded to the nearest integer value [5].

The benefit of using the λ-domain model is that, λ is the slope of the operational convex rate-distortion curve. So, there is a one-to-one correspondence between rate value and λ. Consequently, the λ parameter can be determined in a specific range according to the target bitrate, without interference from the rest of the coding parameters. In addition, adjusting λ parameter can be precise enough since it can take any continuous positive value. Finally, since λ is a parameter that is not needed by the decoder, it shouldn't be sent to the receiver side. Therefore, the higher precision of λ will not raise the bitrate [5].

After specifying the QP value using initial bitrate and (2) and (3), the QP value of the other tiles of each layer should be extracted. Since the various tiles of each layer are captured from a common scene from various viewpoints, they may be similar to each other. Our proposed approach uses this similarity in order to select the QP value of the other tiles of each layer. For this purpose, we generalize the approach suggested for view scalability in [35] to tile-based scalability as follows. As we explained in the features of tile-base scalability, the tiles inside each layer can be predicted from each other using inter-view prediction.

Hence, if the disparity between the first and second tiles of each view, i.e., Tile #12, and Tile #22 in Fig. 2, is low, the Tile #22 can be predicted better from Tile #21 and consequently can be compressed more efficiently. So the bitrate of this tile can be much lower without affecting the overall quality. Hence, the inter-view disparities among various tiles of each layer have a direct impact on prediction efficiency and the total rate of tiles inside the layer.

As such, we generalize the method proposed in [35] and use the concept of inter-view disparity to find the appropriate relationship between the quantization parameters of various tiles using the following equation:

$$QP_{Tn} = QP_{T0} + \frac{1}{inter{-}view\ disparity\ between\ Tile\#0\ and\ Tile\#n} \quad (4)$$

where QP_{T0} is the QP value of the first tile in each layer that is extracted from the previous step and QP_{Tn} is the QP value of tile #n in that layer.

Results and Discussion

In this section, several experiments were performed to show the efficiency of our proposed scalable framework. The results show the effectiveness of the proposed tile-based scalability in terms of computational complexity and the effectiveness of our method in allocating reasonable rate to each base and enhancement layer and their corresponding tiles. Results have been obtained using the MV-HEVC reference software [36]. Table 1 summarizes the properties of our test sequences [37]-[39].

Table 1: Properties of the sequences

Sequence	Original Resolution	Number of Views
Lincoln	2048 × 2048	107
Pantomime	1280 × 960	80
Tower	1280 × 960	80
Vassar	640 × 480	7

For each sequence, we have used three views for our experiments. Each view is partitioned into two or three different tiles. Then, we have considered one base layer and one or two corresponding enhancement layers. Table 2 shows the detail of layers and tiles selection for our test sequences. For instance, for the Lincoln sequence, we have used three views for our experiments, view#00, view#40, and view#60. Each view partitioned to three different tiles, the left tile, the middle tile, and the right tile. Then, we have considered one base layer and two corresponding enhancement layers. The middle tiles are placed in the base layer, and the other ones are placed in enhancement layers, respectively.

It should be noted that, we tried to have selected the views as far away as possible in each video sequence to cover the wide-ranging field of view at the receiver side. Then, the base and enhancement layers are selected based on the ROI parts of the video. The more attractive parts are considered as the base layer.

For instance, if the important objects of the scene are located in the right part of the scene, the tiles of this part are selected as the base layer. Fig. 3 shows the base and enhancement layers of the Tower sequence for more clarification. As we can see, the most important part of the scene is assigned to the base layer and the other parts are assigned to the enhancement layer.

Table 2: Base and enhancement layers and tiles selection for our test sequences

Seq.	View#	Number of layers	Base layer tiles	Enh. layer tiles
Lincoln	3	View#00 View#40 View#60	Middle tiles	Left tiles Right tiles
Pantomime	3	View#00 View#30 View#50	Left tiles	Right tiles
Tower	3	View#30 View#35 View#55	Right tiles	Left tiles
Vassar	3	View#00 View#04 View#07	Right tiles	Left tiles

We also have used the Bjøntegaard-Delta bitrate (BD-bitrate) and Bjøntegaard-PSNR (BD-PSNR) as the measure for Rate-Distortion (RD) performance [40]. BD-bitrate indicates the average difference in bitrate for the same the quality evaluation in PSNR, and BD-PSNR shows the average PSNR difference in dB over the whole range of bitrates.

Fig. 3: The enhancement (left) and base (right) layers of Tower sequence in Tile-based scalability.

In the reported results, negative BD-bitrate means bitrate savings, and positive BD-PSNR means the bitrate increase compared to the anchor case and positive BD-PSNR means the quality improvement over the anchor case.

For the anchor case, we have used the “current” methods that use the λ-domain rate control [5] to find the QP values for all layers and corresponding tiles.

The anchor method extracts the proper QP values of each tile of base and enhancement layers using λ-domain rate control method as follows. The proper QP value for each tile should be extracted using (3). Hence, first, the λ parameter for each tile should be exploited using (2) and the total bitrate of each tile. As discussed before, the total bitrate of the base layer should be assigned according to the minimum bandwidth requirements of all receivers. Since the base layer is the most important part of the scene, the highest portion of the total bitrate should be assigned to this layer. Without loss of generality, we assume that the 2/3 and 1/3 of the total bitrate are assigned to the base and the enhancement layers, respectively.

We have considered four initial QP values, 30, 25, 20, and 15 and encode all of our test sequences using these QP values. Then, 2/3 and 1/3 of the total extracted bitrate are assigned to base and enhancement layers, as suggested.

Since, we have encoded three views of each video sequence, each base and enhancement layer has three tiles. So, the bitrate of each tile is the total bitrate of each layer divided by 3.

In addition, in order to use (2), the α and β parameters

should be exploited. As we have discussed before, these parameters are related to the content of the video. We pre-encoded our test sequences with four different initial QP values, 30, 25, 20, and 15. Then we have used these initial QP values and the total extracted bitrate in (2) to find α and β parameters.

Finally, using the extracted α and β parameters and the bitrate of each tile, the λ value of each tile can be extracted and this λ value can be used in (3) to find the proper QP of each layer.

For our proposed approach, we have used the λ-domain rate control method to find the proper QP value just for the first tile of each layer. The QP values of the other tiles are extracted using the concept of inter-view disparity as explained before. To calculate the inter-view disparity between two tiles, a step which is needed for QP value extraction in (4), we have used the full search approach to estimate disparity between the tiles, accurately.

Table 3 shows the inter-view disparity values for our test sequences measured by this method. As mentioned before, we have considered two enhancement layers for Lincoln sequence and just one enhancement layer for the other ones.

Table 3: The inter-view disparity between Tile0 and the other tiles of each layer for our test sequences

Seq.	Base layer		Enh. layer1		Enh. Layer2	
Lincoln	1.61	1.022	0.60	0.55	0.55	3.06
Pantomime	18.05	18.10	0.09	0.04		
Tower	0.61	5.22	0.48	21.29		
Vassar	0.26	0.52	1	0.65		

We have calculated the QP values of each tile in base and enhancement layers according to (4) and using the mentioned inter-view disparity estimation. The corresponding extracted QP values for the anchor and the proposed method for initial QP=15 is illustrated in Table 4. As the same way, the QP are exploited for anchor and proposed method for initial QP, 20, 25 and 30.

At the next step, each tile of the base and enhancement layers has been coded by the extracted QP values for the anchor and the proposed method using MV-HEVC.

Table 5 provides the coding performance analysis of the proposed approach against the anchor method, where the negative BD-bitrate means bitrate savings compared to the anchor method. As we can see, our method can achieve mush better performance compared to the anchor method in terms of bitrate saving and quality.

Table 4: The extracted QP values for anchor and proposed method for our test sequences for initial QP=15

	Seq.		Lincoln	Pantomime	Tower	Vassar
			QP			
Anchor Method	Base Layer	Tile 0	14	15	15	15
		Tile 1	15	15	15	15
		Tile 2	15	15	15	15
	Enh. Layer1	Tile 0	16	15	15	16
		Tile 1	16	15	15	16
		Tile2	16	15	15	16
	Enh. Layer2	Tile 0	16			
		Tile 1	16			
		Tile 2	16			
Proposed method	Base Layer	Tile 0	14	15	15	15
		Tile 1	15	15	17	20
		Tile 2	15	15	15	17
	Enh. Layer1	Tile 0	16	15	15	16
		Tile 1	18	26	17	51
		Tile2	18	44	15	18
	Enh. Layer2	Tile 0	16			
		Tile 1	18			
		Tile 2	16			

Table 5: Coding performance comparison of the proposed Method against anchor method

Seq.	Base layer		Enh. Layers		Base + Enh. Layers	
	BD-rate	BD-PSNR	BD-rate	BD-PSNR	BD-rate	BD-PSNR
Lincoln	-0.1	0.1	-0.3	0.1	0.2	-0.1
Pantomime	0	0	-30.2	3.7	-26.4	2.3
Tower	-22.7	1.03	-2.7	0.2	-22.6	1.3
Vassar	-1.2	0.3	-24.3	1.83	-21.8	1.3
Avg.	-6	0.3	-14.4	1.5	-17.7	1.2

The corresponding RD curves for the total base and enhancement layers bitrate and PSNR for the Vassar sequence and for the anchor and the proposed method are illustrated in Fig. 4 for visual clarification. As seen, the bitrate is reduced significantly using our proposed method.

We have compared the results of the proposed method with papers as follows.

In [1] a regional bit allocation and rate-distortion optimization method for multiview with depth video coding is proposed that allocate more bits color texture area corresponding depth region and fewer bits to the

color smooth area corresponding depth region. The results of Table 6 compare the coding performance of our proposed method and the method proposed in [41] over the same anchor method where negative BD-bitrate means bitrate savings with respect to the anchor method. As we can see, our method can achieve much better performance compared to the anchor method in terms of bitrate saving and quality.

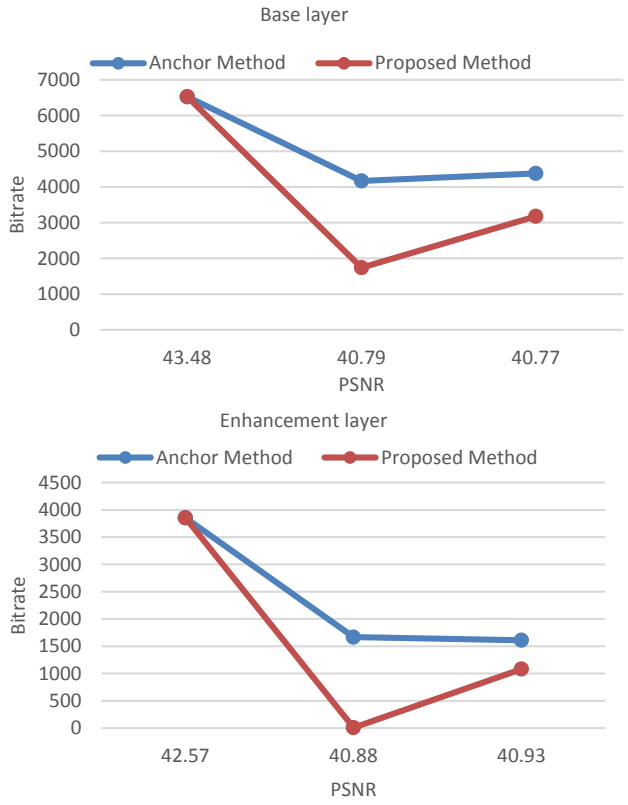


Fig. 4: Rate-Distortion curves for base and enhancement layers for anchor and proposed method of Vassar sequence.

Table 6: The comparison of coding performance of our proposed method against the method proposed in [41]

	Our proposed Method		The method proposed in [41]	
	BD-rate	BD-PSNR	BD-rate	BD-PSNR
Tower	-22.58%	1.35 dB	-19.78%	0.82 dB
Pantomime	-26.40%	2.29 dB	-21.98%	0.19 dB

In [42] a bit allocation optimization method for Multiview Video Coding (MVC) based on stereoscopic visual attention that exploits the visual redundancies from human perceptions. The results of Table 7 compare the coding performance of our proposed method and the method proposed in [2] over the same anchor method where negative BD-bitrate means bitrate savings and negative BD-PSNR shows the performance degradation with respect to the anchor method. The results show that our method can achieve much better performance compared to the anchor method in terms of bitrate saving and quality.

Table 7: The comparison of coding performance of our proposed method against the method proposed in [42]

	Our proposed Method		The method proposed in [42]	
	BD-rate	BD-PSNR	BD-rate	BD-PSNR
Tower	-8.58%	-0.40 dB	-19.78%	0.82 dB
Pantomime	-8.19%	-0.14 dB	-21.98%	0.19 dB

The main advantage of our proposed scalable modality over the previous ones is the low computational complexity of our method. In tile-based scalability, the tiles of base and enhancement layers can be coded independently from each other in parallel. It can improve the total computational time of our method over the previous scalable modalities such as view scalability [35].

We have implemented view scalability for our test sequences and coded the data of base end enhancement layers in both tile-based scalability and view scalability. Table 8 shows the processing time of our proposed algorithm compared to the method proposed in [35] which uses view scalability that cannot benefit from parallel processing. Both methods are run on an Intel i7-4790 CPU @ 3.6 GHz. As we can see, the processing time of our proposed that can uses parallel processing is much better.

Table 8: Comparison between the computational complexity of tile-based scalability and view scalability [35]

Sequences	Processing time of tile-base scalability (sec)	Processing time of view scalability (sec)
Lincoln	3240.267	6007
Pantomime	156.6	256.99
Tower	199.93	242.14
Vassar	50.3	103
Average	911.77	1652.28

Conclusion

In this paper, we presented a novel scalable framework for free-viewpoint video application. This framework proposes new tile-based scalability to assign proper regions to base and enhancement layers according to the features of free viewpoint scalability.

Then, a rate control approach is proposed to assign an appropriate bitrate to base and enhancement layers.

Besides, using tile coding may lead to reduce computational complexity. Experimental results showed that the proposed method could achieve a good coding efficiency over the anchor method that used the λ-domain rate control method with much less computational complexity.

Author Contributions

H. Roodaki, designed the experiments, collected the data, carried out the data analysis, interpreted the results, and wrote the manuscript.

Acknowledgment

The author received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

There is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Abbreviations

3DTV	3D television
MVC	Multiview video coding
SMVC	Scalable multiview video coding
MV-HEVC	Multiview video-High efficiency video coding
QP	Quantization parameter

References

[1] C. C. Lee, A. Tabatabai, K. Tashiro, “Free viewpoint video (FVV) survey and future research direction,” APSIPA Trans. Signal Inf. Process., 4: 1-10, 2015.

[2] Y. Zheng, J. Feng, H. Ma, Y. Chen, “H.264 ROI coding based on visual perception,” in Proc. 5th International Conference on Visual Information Engineering. Xian China, China: 829-834, 2008.

[3] J. Zhang, L. Zhuo, Y. Zhao, “Region of interest detection based on visual perception model,” Int. J. Pattern Recognit. Artif. Intell., 26(02), 2012.

[4] G. J Sullivan, J Ohm, W. J. Han, T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” IEEE Trans. Circuits Syst. Video Technol., 22(12): 1649-1668, 2012.

[5] B. Li, H. Li, L. Li, J. Zhang, ”λ domain rate control algorithm for high efficiency video coding,” IEEE Trans. Image Process., 23(9): 3841-3854, 2014.

[6] H. Schwarz, D. Marpe, T. Wiegand, “Overview of the scalable video coding extension of the H.264/AVC standard,” IEEE Trans. Circuits Syst. Video Technol., 17(9): 1103-1120, 2007.

[7] D. Grois, E. Kaminsky, O. Hadar, “Dynamically adjustable and scalable ROI video coding,” in Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB): 1-5, 2010.

[8] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, Y. Yashima, “View scalable multiview video coding using 3-D warping with depth map,” IEEE Trans. Circuits Syst. Video Technol. 17(11): 1485-1495, 2007.

[9] H. Yo Sung, O. Kwan Jung, “Overview of multi-view video coding. 14th International Workshop on Systems Signals and Image,” in Proc. 6th EURASIP Conference focused on Speech and Image Processing Multimedia Communications and Services, 2007.

[10] S. Shimizu, H. Kimata, K. Kamikura, Y. Yashima, “Free-viewpoint scalable multi-view video coding using panoramic mosaic depth maps,” in Proc. 16th European Signal Processing Conference, 2008.

[11] H. Roodaki, S. Shirmohammadi, “Scalable multiview video coding for immersive video streaming systems,” in Proc. Visual Communications and Image Processing (VCIP), 2016.

[12] N. Ozbek, A. Murat Tekalp, E. Turhan Tunali, “A new scalable multi-view video coding configuration for robust selective streaming of free-viewpoint TV,” in Proc. IEEE International Conference on Multimedia and Expo, Beijing: 1155-1158, 2007.

[13] T. Yan, I. H. Ra, Q. Zhang, H. Xu, L. Huang, “A novel rate control algorithm based on ρ model for multiview high efficiency video coding, “Electronics, 9(1): 166, 2020.

[14] T. Yan, I. H. Ra, D. Liu, D. Chen, Y. Youhao, S. Hou, “Rate control based on similarity analysis in multi-view video coding,” in Proc. 9th International Conference on Information Technology Convergence and Services (ITCSE 2020), 2020.

[15] T. Yan, I. H. Ra, Q. Zhang, H. Wen, H. Xu, S. Chen, “Rate control algorithm for multiview video coding based on human visual characteristics,” Int. J. Performability Eng., 14(8): 1913-1921, 2018.

[16] T. Yan, I. H. Ra, H. Wen, M. H. Weng, Q. Zhang, Y. Chen, “CTU layer rate control algorithm in scene change video for free-viewpoint video,” IEEE Access, 8: 24549-24560, 2020.

[17] T. Yan, P. An, L. Shen, Q. Zhang, Z. Zhang, “Rate control algorithm for multi-view video coding based on correlation analysis,” in Proc. Symposium on Photonics and Optoelectronics: 1-4, 2009.

[18] L. Li, B. Li, D. Liu, H. Li “λ-domain rate control algorithm for HEVC scalable extension,” IEEE Trans. Multimedia, 18(10): 2023-2039, 2016.

[19] P. J. Lee, Y. C. Lai, “Vision perceptual based rate control algorithm for multi-view video coding,” in Proc. International conference on system science and engineering (ICSSE), 2011.

[20] S. Park, D. Sim, “An efficient rate-control algorithm for multi-view video coding,” in Proc. IEEE 13th International Symposium on Consumer Electronics: 115-118, 2009.

[21] A. Fiengo, G. Chierchia, M. Cagnazzo, B. Pesquet-Popescu, “Convex optimization for frame-level rate allocation in MV-HEVC,” in Proc. IEEE International Conference on Image Processing (ICIP), 2016.

[22] M. Cordina, C.J. Debono, “A depth map rate control algorithm for HEVC Multi-View Video plus Depth,” in proc. IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2016.

[23] R. Abolfathi, H. Roodaki, S. Shirmohammadi, "A novel rate control method for free-viewpoint video in MV-HEVC," in Proc. 2019 International Conference on Computing, Networking and Communications (ICNC): 582-587, 2019.

[24] T. Yan, I. Ra, H. Wen, M. Weng, Q. Zhang, Y. Che, "CTU layer rate control algorithm in scene change video for free-viewpoint video," IEEE Access, 8: 24549-24560, 2020.

[25] J. Kilner, J. Starck, A. Hilton, “A comparative study of free-viewpoint video techniques for sports events,” in proc. 3rd European Conference on Visual Media Production, 2006.

[26] C. M. Privitera, L. W. Stark, “Algorithms for defining visual regions-of-interest: comparison with eye fixations,” IEEE Trans. Pattern Anal. Mach. Intell., 22(9): 970-982, 2000.

[27] J. L. Solka, D. J. Marchette, B. C. Wallet, V. L Irwin, G. W. Rogers, “Identification of man-made regions in unmanned aerial vehicle imagery and videos,” IEEE Trans. Pattern Anal. Mach. Intell., 20(8): 852-857, 1998.

[28] A. Mohan, C. Papageorgiou, T. Poggio, "Example-based object detection in images by components," IEEE Trans. Pattern Anal. Mach. Intell., 23(4): 349-361, 2001.

[29] T. M. Stough, C. E. Brodley, "Focusing attention on objects of interest using multiple matched filters," IEEE Trans. Image Process., 10(3): 419-426, 2001.

[30] K. Misra, A. Segall, M. Horowitz, S. Xu, A. Fuldseth, M. Zhou, "An overview of tiles in HEVC," IEEE J. Sel. Top. Signal Process., 7(6): 969-977, 2013.

[31] G. Corrêa, P. Assunção, L. Agostini, A. L. da Silva Cruz, "Performance and computational complexity assessment of high-efficiency video encoders," IEEE Trans. Circuits Syst. Video Technol., 22(12): 1899- 1909, 2012.

[32] M. Viitanen, J. Vanne, T. D. Hämäläinen, M. Gabbouj, J. Lainema, "Complexity analysis of next-generation HEVC decoder," in Proc. IEEE International Symposium on Circuits and Systems: 882-885, 2012.

[33] A. Fuldseth, M. Horowitz, S. Xu, K. Misra, A. Segall, M. Zhou, "Tiles for managing computational complexity of video encoding and decoding," in Proc. Picture Coding Symposium (PCS): 389-392, 2012.

[34] J. Wen, M. Fang, M. Tang, K. Wu, "R-(lambda) model based improved rate control for HEVC with Pre-Encoding," in Proc. Data Compression Conference (DCC): 53-62, 2015.

[35] H. Roodaki, M. R. Hashemi, S. Shirmohammadi, "Rate-distortion optimization for scalable multi-view video coding," in Proc. IEEE International Conference on Multimedia and Expo (ICME), 2014.

[36] G. Sullivan, J. M. Boyce, Y. Chen, J. R. Ohm, C. A. Segall, A. Vetro, "Standardized extensions of high efficiency video coding (HEVC)," IEEE J. Sel. Top. Signal Process., 7(6): 1001-1016, 2013.

[37] Tanimoto Laboratory, http://www.fujii.nuee.nagoya-u.ac.jp/~fukushima/mpegftv/, last ACCESS on. March, 2021.

[38] Merl, http://www.merl.com/pub/avetro/mvc-testseq/orig-yuv/vassar/, last ACCESS on March, 2021.

[39] ftp://ftp.research.microsoft.com/users/fvv/, last ACCESS on March, 2021.

[40] G. Bjøntegaard, "Calculation of average PSNR differences between RD curves," ITU T SG16/Q6, Doc. VCEG-M33, 2001.

[41] Y. Zhang, S. Kwong, L. Xu, S. Hu, G. Jiang , C. J. Kuo, "Regional bit allocation and rate distortion optimization for multiview depth video coding with view synthesis distortion model," IEEE Trans. Image Process., 22(9): 3497-3512, 2013.

[42] Y. Zhang, G. Jiang, M. Yu, et al., "Stereoscopic visual attention-based regional bit allocation optimization for multiview video coding," EURASIP J. Adv. Signal Process. 848713, 2010.

Biographies

Hoda Roodaki was born in Tehran, Iran in 1982. She received the B.S. and M.S. degrees in Computer engineering from the University of Tehran, and Sharif University of Technology, in 2005 and 2007 and the Ph.D. degree in Computer Architecture from university of Tehran, Iran in 2014. Since 2015, she has been an Assistant Professor with the Computer Engineering Department, K. N. Toosi University, Tehran, Iran. Her research interests include Multi-view/3D and 360-degree video coding, Scalable video coding, video quality assessment and cloud gaming.

- Email: hroodaki@kntu.ac.ir
- ORCID: 0000-0002-3575-0587
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://wp.kntu.ac.ir/hroodaki/

How to cite this paper:
H. Roodaki, "An efficient Region-of-Interest (ROI) based scalable framework for free viewpoint video application," J. Electr. Comput. Eng. Innovations, 12(1): 283-293, 2024.

DOI: 10.22061/jecei.2023.7934.455

URL: https://jecei.sru.ac.ir/article_2017.html