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

Research paper

A Variational Level Set Approach to Multiphase Multi-Object Tracking in Camera Network Base on Deep Features

***E. Pazouki*[1,*], *M. Rahmati*[2]**

[1]*Artificial Intelligence Department, Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.*

[2]*Artificial Intelligent and Robotics Department, Faculty of Computer and Information Technology Engineering, Amirkabir University of Technology, Tehran, Iran.*

Article Info

Article History:
Received 02 September 2020
Reviewed 15 December 2020
Revised 25 January 2021
Accepted 01 March 2021

Keywords:
Multi-object tracking
Camera network tracking
MultiPhase level set representation
Variational tracking
Deep features

*Corresponding Author's Email Address:
ehsan.pazouki@sru.ac.ir

Abstract

Background and Objectives: Object tracking in video streams is one of the issues in machine vision that has many applications. Depending on the type of the object, the number of objects and other inputs used in tracking, object tracking is divided into several different categories. Multi-object tracking in a camera network is one of the most complex types of object tracking. In this type of tracking, the goal of the algorithm is to extract the persistent trace of several objects moving simultaneously in a wide area that is monitored by a network of cameras. This type of tracking is often done in two steps. In the first step, the traces of each object in each camera is called tracklets are extracted. Then, the persistent trace of the objects are obtained by associating the extracted tracklets of all cameras in the monitored wide area. Here, we introduce a novel variational approach based on the deep features to associate the tracklets.

Methods: For this purpose a variational model with multiphase level set representation is introduced. The persistent trace of all objects are obtained by optimizing the proposed variational model. The proposed variational model is optimized by employing the Euler-Lagrange equation. CNN and deep learning are used to extract the deep features of appearance and motion of objects. Here, a ResNet50 network that is pre-trained on ImageNet and a transformer neural network which is trained with motion parameters of tracklets such as acceleration and orientation change rate are used for extracting deep features.

Results: The multiphase model using deep features presented in this paper provide 9% better results than the multiphase model without deep features based on TCF and FS metrics and 8% better results based on MT metric.

Conclusion: The results on the three well-known datasets which are real and a synthesized dataset show that the proposed model takes competitive performance, while using less extra context information of the camera network and objects, compared to the other proposed methods. The evaluations show the quality of the proposed model in solving complex problems using the minimum required initial knowledge.

Introduction

Object tracking in wide areas which is monitored by a network of cameras is one of the most challenging issues in the machine vision. The tracking module extract

trajectories of objects presented and which are viewed by a camera network covering the area. The high level analysis is performed on the extracted trajectories and appropriate alarms are set. Multi-object tracking in a camera network is performed in two steps [1]. In the first step, moving objects are captured by all of the cameras within the view of camera network and are tracked by a tracking algorithms. Extensive research has been reported to deal with single camera tracking [2], [3]. There are various challenges exist for extracting accurate trace of objects in each camera, among them pose and illumination variations, occlusion, clutter and sensor noise are more common. Advancement of algorithms to resolve object tracking in a single camera have reached to a satisfactory extent [4], [5]. In this paper, the admissible trace of the objects in a single camera using the existing algorithms are extracted and are called tracklets. In the second step of multi-object tracking of a camera network, the extracted tracklets are associated to the corresponding objects and the persistent trace of objects are extracted. Usually, the association is performed based on the appearance of objects and motion models of the objects which are extracted from the tracklets. The solution space of the association task grows exponentially with the number of the tracklets [6]. Our proposed method provides a remedy for this problem. This problem is an ill-posed inverse problem [7]. The tracklets are the observation and are assumed to be known and the persistent trace is the ideal output and unknown. The variational model is an effective solution for solving the ill-posed inverse problem [8]. This model solves the ill-posed inverse problem in the image processing and computer vision [9], [10], [11] and [12]. In this paper, our main objective is to propose a variational multiphase model for solving the association problem.

Previous attempts on associating tracklets with corresponding objects place different restricted assumptions on the problem, thus deviating their application from real world problem [1]. In [13], [14] and [15], first order Markovian model is used for association. Using Markovian model for tracking objects in a single camera is a reasonable assumption but in a non-overlapped camera network is not a good choice where this assumption increases the probability of a wrong association. In some previous researches [14] and [16], the topology of cameras and the moving model of the objects are assumed to be known, while it is difficult to extract. Also, modeling the moving pattern and updating the changes are challenging.

Multi-object tracking in a camera networks is important issue. Associating of corresponding objects between different cameras distinguishes single camera and multi-camera tracking. Thus, the main challenge of the proposed model for tracking multi-object in camera networks is the association [1]. In some studies [17], [13], [16], [14], [18], and [6], the main goal is to associate the traces of objects and extract their persistent trace. In these researches, various tools and algorithms are used, including a planar tracking correspondence model (TCM), a Bayesian modeling, Parzen windows, the path coverage of a directed graph, a multi-objective optimization framework, a path smoothness, a statistical model function, a graphical representation and a network flow algorithm for modeling and association. In [19], an approach to improve the detection and tracking performance in multi-camera scenarios with overlapping field-of-views is provided, which allows for better handling of occlusion problem. It mainly includes monocular people detection, projection, fusion, probabilistic occupancy map generation and multi-object tracking steps. The objective of the study is to detect and track individuals within a designated open area where multiple cameras are set up, implement a robust multi-camera people detection and tracking method and improve the experimental performance.

With the introduction of deep learning [20] methods and the development of its applications in various issues, the use of models based on deep learning in objects tracking was also introduced in different methods. In some methods, the features of deep neural networks are used as visual or temporal features of objects [21] and [22]. For example in [21] the 2048-dim fully connected layer of ResNet50 [23] before the classification layer is used to represent the appearance of the objects. In some methods, models based on convolutional neural network (CNN) and deep learning are used to extract tracklets or persistent trace of the objects [22], [24], [25] and [26]. In [22], a CNN-based model called TrackletNet is introduced that uses the a graph to extract the trace of objects. In [24] a Long Short-Term Memory (LSTM) is used for tracking object. In [25] is provided a model based on Recurrent Neural Network(RNN) that models the changes made in the object by updating the features of the tracking object and performs better tracking of the object. In [26] a CNN-based model called Siam R-CNN is proposed which combine a Siamese re-detection architecture with a tracklet-based dynamic programming algorithm. In [27] a tracklet processing algorithm cleave and re-connect tracklets on crowd or long term occlusion which uses Siamese Bi-Gated Recurrent Unit (GRU). In this study, the tracklet is generated using deep features which are extracted by CNN and RNN to create the high-confidence tracklet candidates in sparse scenario. This proposed neural network model is trained using a dataset which contains more than 95160 pedestrian images. In [28] a multi-object tracking framework called DROP (Deep Re-identification Occlusion Processing) is

proposed. A lightweight convolutional neural network that can solve the re-tracking problem is constructed by increasing and learning the affinity of appearance features of the same object in different frames. In this study the occlusion of the object is judged using the data association result of the appearance features of the object, and to reduce the matching error by improving the data association formula.

In some studies, the main objective is to introduce compositional features that are more effective in recognizing objects. In [29] a procedure is introduced which, using the classic features PHOG and CS-LBP and combining them with deep features and also using a new feature selection tools such as JEKNN, proposes an efficient combined features for classifying objects. This combined feature is able to significantly improve results relative to similar models. In [30] deep features are combined with multiview features and a set of features is obtained for recognizing human behavior, which has good results compared to similar models. In order to model the objects with the aim of properly classifying them in [31], the combination of deep features, Very Deep Convolutional Networks for Large-Scale Image Recognition and Inception V3, are used. The obtained combined features are able to provide good sustainable recognition rate in object classification. Also in [32], by combining the deep features obtained from the two deep networks, VGG and AlexNet, with SIFT which is the classical feature for object detection, a combined feature has been created with the help of Reyni entropy-controlled method, which provides good results in classifying objects.

The combined features, some of which were reviewed in the above studies, are not practical in object tracking applications due to the considerable computational complexity involved in extracting them, as in object tracking the features of all objects in the image must be calculated for each frame. If they are used, the response time of the tracking algorithm will be significantly increased, which is not tolerable and acceptable in many real problems.

Another method of tracking objects is to use variational model, examples of which are presented in [33] and [34].

In [33] a variational model called single phase variational is introduced which is tracked the multi-objects in multi-camera network of wide area surveillance system. In [34] a novel variational model called multi phase variational model is proposed which is used the RGB color histogram as appearance model of the objects and the acceleration and orientation change rate as the motion model of the objects.

In this paper, we proposed a novel deep variational method for associating the tracklets. In this method, persistent trace of each object is represented as a multiphase level set function. By solving this association problem with less restrictive assumption, the optimum solution is reachable and the method is more general and usable in realistic scenarios. CNN and deep learning are used to extract the features of appearance and motion of objects.

In this paper, we propose a multiphase variational model for associating tracklets based on deep features. The proposed model is a variational optimization model that is converted to an Ordinary Differential Equation (ODE) which is solved numerically for extracting the persistent trace of objects.

The structure of the paper is as follows. In Section Variational Model the proposed variational model and solving method are presented. The experimental results of the proposed model on the real and synthesized datasets is given in Section Experimental Results and finally, the conclusions is presented in Section Conclusion.

Variational Model

In this section, the proposed variational model and solving method are presented.

A. Problem Formulation

The main goal is to extract the persistent trace of the objects in a camera network which monitors the wide area. The wide area is monitored with k cameras $C_1, \dots, C_k$. Each image captured by a camera is mapped to the world plane using calibration parameters of the cameras $C_{M_1}, \dots, C_{M_k}$. Unknown n number of objects $P_1, \dots, P_n$ moves in a wide area which is monitored by a network of cameras.

Any object denoted by P_τ that is moving in the area which is covered by each camera in $[t_s, t_e]$ period is tracked by its corresponding single-camera tracking algorithm [2] and [4]. Therefore, for each camera C_i of the camera network, a set of the tracks is called tracklets T_{O_i} is exploited. So, the set of all tracklets of all cameras $T_C = \{T_{O_1}, \dots, T_{O_K}\}$ is obtained. For generating persistent track of the objects R_P, the obtained tracklets T_C are associated and the tracklets of each object in T_C are corresponded and persistent trace of this object is extracted and is denoted as r_{p_x}. By associating the tracklets of objects, the persistent trace of all objects $R_P = \{r_{p_1}, \dots, r_{p_n}\}$ is obtained. Therefore, the problem is represented as,

$$R_P = Tracker(T_C, \{C_{M_1}, \dots, C_{M_k}\}) \quad (1)$$

which it means the proposed $Tracker(\cdot)$ algorithm computes the persistent trace of objects as R_P using the extracted *tracklets* T_C and camera calibration parameters $\{C_{M_1}, \dots, C_{M_k}\}$. In our proposed method, this problem is modeled with a variational optimization

model as follow,

$$J[R_P] = \int_{r_{p_\tau} \in R_P} \left(\lambda_1 \times \int_{T_l \in T_C} CLS_{r_{p_\tau}}(T_l) dl + \lambda_2 \times \int_{T_l \in r_{p_\tau}} SM_{r_{P_\tau}}(T_l) dl \right) d\tau \quad (2)$$

where $CLS_{r_{p_\tau}}(\cdot)$ is *closeness* part of the variational model, $SM_{r_{P_\tau}}(\cdot)$ is *smoothness* part of the model and $\lambda_1 > 0$ and $\lambda_2 > 0$ are positive constants. The clossness control similarity between tracklets and the smoothness control variations between tracklets.

In this paper, the closeness and smoothness parts which are proposed in [34] are used. The appearance model which is used in this paper is the 2048-dim fullyconnected layer before the classification layer of a ResNet50 [23] network that pre-trained on ImageNet to represent appearance of the objects. A transformer neural network [35] which is trained with motion parameters of tracklets such as acceleration and orientation change rate is used as a motion model of objects.

The transformer model which is used in this research is the same as the proposed model in [35] and is used with the same parameters and settings. In this research, two transformer models have been used, one to estimate the change rate of the angle of the object and the other to estimate the change rate of the acceleration of the object. To train both models, all tracklets which are the inputs of the tracking problem are used as the training set.

Also the parameters and algorithms which are used for training are the same as [35]. The angle change rate and accelration change rate of the tracklets with a frequency of 1 Hz for each tracklets is calculated in the form of an array as a time series, and the time series which are created are used to train two transformer models.

When calculating the closeness and smoothness of the variational model, the angle change rate and acceleration change rate are estimated based on the assosiation of tracklets for blind areas between two adjacent tracklets and then used in calculations.

In brief, the proposed model is a variational energy function declared as (2) which the extracted *tracklets* T_C are its inputs and the persistent traces of the objects R_P are their outputs.

In order to use this proposed model, an appropriate representation of the variational energy function must be employed. We propose a multiphase level set representation to solve the optimization model.

B. Multi Phase Level Set Representation

Proposing the level set representation of the model starts with representing persistent trace of the objects as,

$$\begin{cases} T_l \in r_{P_x} & if\ \varphi_x(l) \geq 0 \\ T_l \notin r_{P_x} & if\ \varphi_x(l) < 0 \end{cases} \quad (3)$$

where persistent trace of each object r_{P_x} is presented as a level set function φ_x and persistent trace of all tracked objects are presented as $\phi = \{\varphi_1, \dots, \varphi_b\}$ which is a multiphase level set function. As stated in previous section the number of the objects which move in the wide area n is unknown and the number of tracked objects b is not necessarily equal to n. Now, the (2) can be rewritten as,

$$J[\phi] = \int_{\tau=1}^{|\phi|} \left(\lambda_1 \times \int_{l=1}^{|T_C|} CLS_{\varphi_\tau}(l) dl + \lambda_2 \times \int_{l=1}^{|T_C|} SM_{\varphi_\tau}(l) \times H(\varphi_\tau(l)) dl \right) d\tau \quad (4)$$

where $H(\varphi_\tau(\cdot))$ is the Heaviside function and is defined as,

$$H(Z) = \begin{cases} 1 & Z \geq 0 \\ 0 & Z < 0 \end{cases} \quad (5)$$

The closeness and smoothness equations are redefined based on the level set representation, and are provided in [34].

C. Optimizing the Energy Function

In order to solve the persistent trace problem, the presented level set representation model must be optimized.

For optimizing this model, the Euler-Lagrange equation is used [36]. It must be noted that the regularized versions of the function $H(z)$,denoted here by $H_\epsilon(z)$, is used. The regularized version of (4) is defined as,

$$J_\epsilon[\phi] = \int_{\tau=1}^{|\phi|} \left(\lambda_1 \times \int_{l=1}^{|T_C|} CLS_{\varphi_\tau,\epsilon}(l) dl + \lambda_2 \times \int_{l=1}^{|T_C|} SM_{\varphi_\tau,\epsilon}(l) \times H_\epsilon(\varphi_\tau(l)) dl \right) d\tau \quad (6)$$

where regularized version of the closeness and smoothness part of the model are redefined in [34].

In this paper, the regularized version of $H(z)$ which is proposed in [37] is used. this function is defined as,

$$H_\epsilon(z) = \frac{1}{2}\left(1 + \frac{2}{\pi}\arctan\left(\frac{z}{\epsilon}\right)\right) \quad (7)$$

The persistent trace of the objects is obtained by solving following optimization problem,

$$\phi^* = \underset{\phi}{argmin} J_\epsilon[\phi]. \quad (8)$$

This optimization problem is solved by fixing, $F_{A,\epsilon}$, $F_{M,\epsilon}$, $MeanF_{A,\epsilon}$ and $MeanF_{M,\epsilon}$ then $J_\epsilon[\phi]$ is minimized with respect to ϕ using Euler-Lagrange equation [7]. The descent direction is parameterized by an artificial time

$t > 0$. Then, the equation $\varphi_\tau(t, l)$ is optimized. The numerical procedure required to solve (8) is presented in [34].

Experimental Results

The performance of our proposed model is evaluated by performing several experiments using three challenging real datasets and one synthesized complex dataset.

The quality of the results of our experiments are determined based on well-known metrics used in camera network applications. The datasets and the metrics that are used are introduced in the following subsections.

A. Datasets

In this paper four challenging datasets are used. Three of them are real video sequences and another is synthesized data which is developed in our Lab.

First, the CAVIAR dataset [38] and [39] is collected in a shopping mall corridor with two cameras.

Second, the NGSIM dataset [40] is captured from Peachtree street located in Atlanta, Georgia by using eight synchronized cameras.

Third, the PETS2009 dataset [41] was collected through 8 cameras which are set up to monitor a road corner of a university campus. But, in this paper only four cameras are used.

In order to generate thorough annotated dataset according to every complicated surveillance scenario based on the given information a tool has been developed [42] for syntesing vitual data.

Here, a synthesized dataset with 6 cameras and 10 objects is used as forth experimental dataset.

The characteristics of these dataset are presented in Table 1.

The datasets' details and the way they are used in this paper are consistent with the experiments performed in [34]. In order to show the status of the datasets, their figures are provided follow.

In Fig. 1 the sample image of one camera of the CAVIAR is presented and some samples of objects in different pose of this dataset are shown in Fig. 2.

In Fig. 3 five sample image of five different cameras of the NGSIM dataset are shown.

In Fig. 4 the sample images of two cars for 3 different pose are presented.

In Fig. 5 the sample images of the four different cameras of PETS2009 dataset are shown and the images of two different objects from different poses are shown in Fig. 6.

In Fig. 7 the world plane image of a wide area covered by 6 cameras from the simulator is presented.

In Fig. 8 some sample images of the objects from simulator are shown.

Fig. 1: A sample image from one of the cameras related to the CAVIAR dataset.

B. Evaluation Metrics

For evaluating the proposed model quantitatively, 10 well known metrics which are commonly used in scope of the proposed model are selected and the quality of the proposed model is measured using these metrics. The nomenclature and some information about these metrics are given in Table 2. Also, details of these metrics are provided in [34].

Fig. 2: Different pose of persons in the CAVIAR: a) First person (First Pose); b) First person (Second Pose); c) First person (Third Pose) ;d) Second person (First Pose); e) Second person (Second Pose); f) Second person (Third Pose).

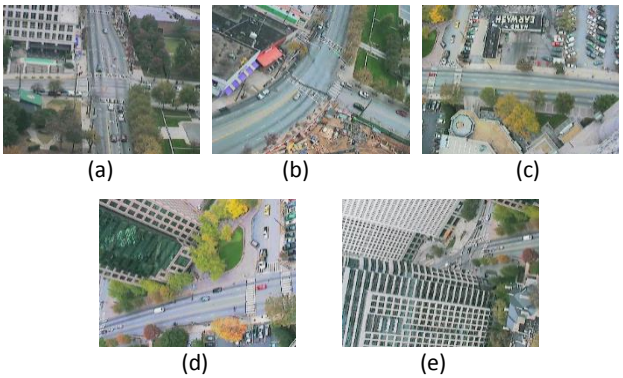
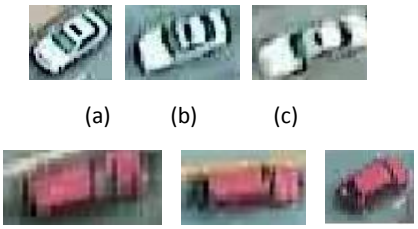


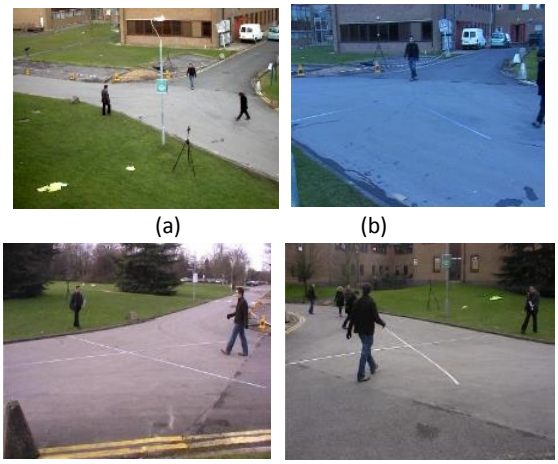
Fig. 3: The sample image of five cameras of the NGSIM; a) First camera; b) Second camera; c) Third camera; d) Fourth camera; e) Fifth camera.



(a) (b) (c)

(d) (e) (f)

Fig. 4: Different pose of cars in the NGSIM: a) First car (First Pose); b) First car (Second Pose); c) First car (Third Pose); d) Second car (First Pose); e) Second car (Second Pose); f) Second car (Third Pose).



(a) (b)

(c) (d)

Fig. 5: The sample image of four cameras of the PETS2009; a) First camera; b) Second camera; c) Third camera; d) Fourth camera.

(a) (b) (c)

(d) (e) (f)

Fig. 6: Different pose of persons in the PETS2009: a) First person(First Pose); b) First person(Second Pose); c) First person (Third Pose); d) Second person (First Pose); e) Second person (Second Pose); f) Second person (Third Pose).

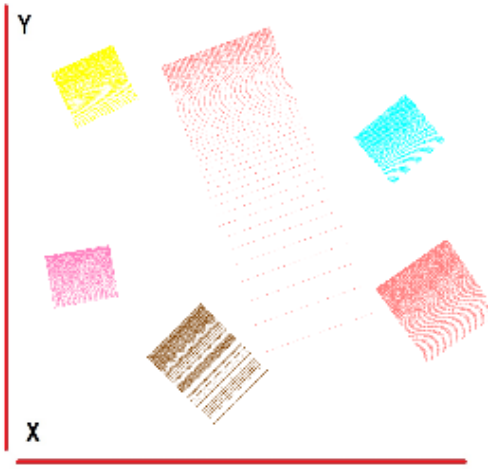


Fig. 7: The top view of the wide area of the synthesized dataset with coverage of cameras.

(a) (b) (c)

(d) (e) (f)

Fig. 8: Different pose of persons in the syntesized dataset: a) First person (First Pose); b) First person (Second Pose); c) First person (Third Pose); d) Second person (First Pose); e) Second person (Second Pose); f) Second person (Third Pose).

C. Results and Discussion

For evaluating the performance of the proposed model, the results are compared with four similar models in Table 3 and Table 4. As provided in Table 3 the proposed model tracks objects in wide area with the

average TCF metric of 78.8% and FS metric of 79.26% which means it extracts more than 79% of the objects' persistent trace. These results show that the multiphase model using deep features presented in this paper can provide 9% better results than the multiphase model without deep features [34].

Also, the proposed model tracks objects with average MT metric of 85.60% which means it tracks more than 80% of the persistent trace of more than 85% of the objects.

As a result, the deep features in terms of MT metric improve by an average of 8% compared to the other features [34]. In order to better compare the results of the proposed model with other models, two charts are presented in Fig. 9 and Fig. 10.

Figure 9 presents a chart for showing the effect of using the variational model and deep features simultaneously compared to the variational based models without deep features.

The numbers in this chart are obtained by averaging the results obtained from the entire datasets.

As can be seen in this figure, the proposed model in most metrics has provided better results than both variational models without deep features.

Also, in Fig. 10 the proposed model is compared with the results of the two other methods that are not variational based, which show the better performance of the proposed model. In other words, the proposed model presents competitive results compare to the similar models.

In Fig. 11, the computed persistent trace of one of the objects of NGSIM dataset and ground truth persistent trace of this object in the world plane are shown. As illustrated in this figure, the object are tracked in three cameras of the camera network. In Fig. 12, the persistent trace result of an object of PETS2009 dataset in two cameras is presented. Also, the extracted persistent trace of an object of synthesized dataset which has been tracked in four cameras of the camera network is given in Fig. 13.

Table 1: The characteristics of the datasets

Name	Wide Area Width(meter)	Wide Area Height(meter)	#Cameras	#Tracklets	#objects
CAVIAR	30	65	1	413	88
NGSIM	150	650	5	691	195
PETS	50	55	4	53	10
Synthesis	50	50	6	102	10

Table 2: The characteristics of the Metrics

Name	Abbreviation	Unit	Minimum	Maximum	Goal
Track Completion Factor [18]	TCF	Percent%	0%	100%	Max
Track Fragmentation [18]	TF	Numerical#	1	-	Min
Physical Object ID Fragmentation [43]	POIF	Numerical#	0	1	Max
Precision [43]	PT	Percent%	0%	100%	Max
Sensitivity [43]	ST	Percent%	0%	100%	Max
F-Score [43]	FS	Percent%	0%	100%	Max
ID Switching [44]	IDS	Numerical#	0	-	Min
Fragment [44]	FG	Numerical#	0	-	Min
Mostly Tracked [44]	MT	Percent%	0%	100%	Max
Mostly Lost [44]	ML	Percent%	0%	100%	Min

Table 3: The results of the proposed model

Metrics	Our Proposed Model				B. Song [39]	R. Pless [18]
	CAVIAR	NGSIM	PETS2009	Synthesized	CAVIAR	NGSIM
TCF	73.14%	79.6%	77.26%	85.18%	-	67%
TF	1.52	1.18	1.16	1.28	-	1.39
POIF	0.38	0.34	0.48	0.35	-	-
PT	84.21%	91.27%	76.12%	87.74%	-	-
ST	73.28%	70.11%	77.29%	78.15%	-	-
FS	78.37%	79.3%	76.7%	82.67%	-	-
IDS	7	3	4	5	8	-
FG	5	8	4	8	6	-
MT	85.12%	89.39%	84.86%	83.02%	84.0%	-
ML	0%	0%	0%	0%	4.0%	-

Table 4: The Results of Single Variational Model [33] and Multi Variational Model [34]

Metrics	CAVIAR		NGSIM		PETS2009		Synthesized	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi
TCF	71%	71.97%	74%	76.3%	74%	75.52%	83%	64.13%
TF	1.61	2	1.21	1.27	2	1.22	2	2.5
POIF	0.37	0.31	0.31	0.32	0.46	0.41	0.33	0.28
PT	83.76%	79.68%	80.43%	89.39%	75.35%	70.04%	87.50%	76.52%
ST	72.58%	68.56%	69.83%	62.12%	76.14%	74.62%	77.62%	68.85%
FS	75.74%	71.33%	71.09%	69.69%	72.43%	68.79%	79.61%	69.40%
IDS	9	11	28	3	5	5	8	6
FG	6	7	35	9	8	5	10	12
MT	80.10%	78.32%	73.40%	87.09%	83.33%	70%	81.00%	75%
ML	2.00%	0%	12%	0%	0%	0%	0%	0%

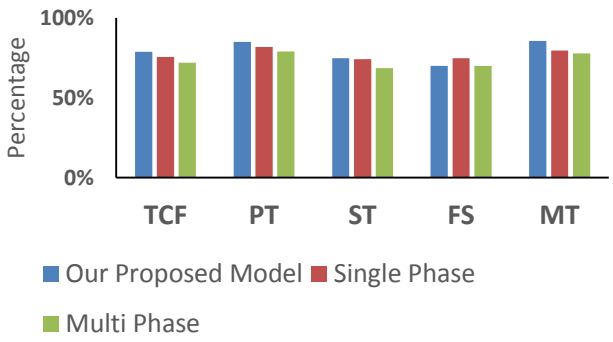


Fig. 9: Graph comparing the results of the proposed model with two other variational models, including the single-phase model [33] and the multi-phase model [34] without using deep features.

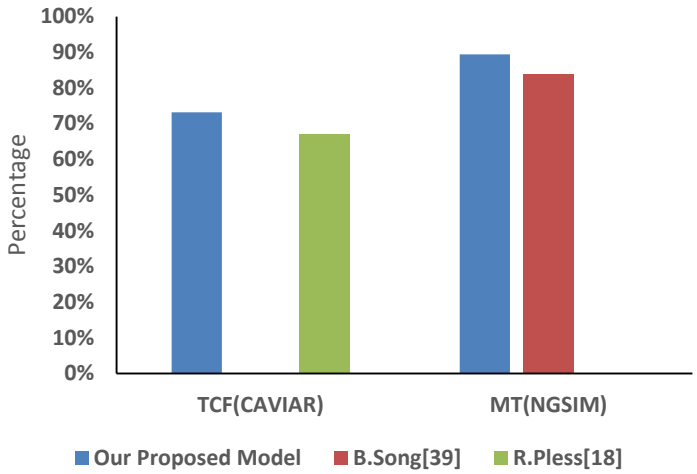
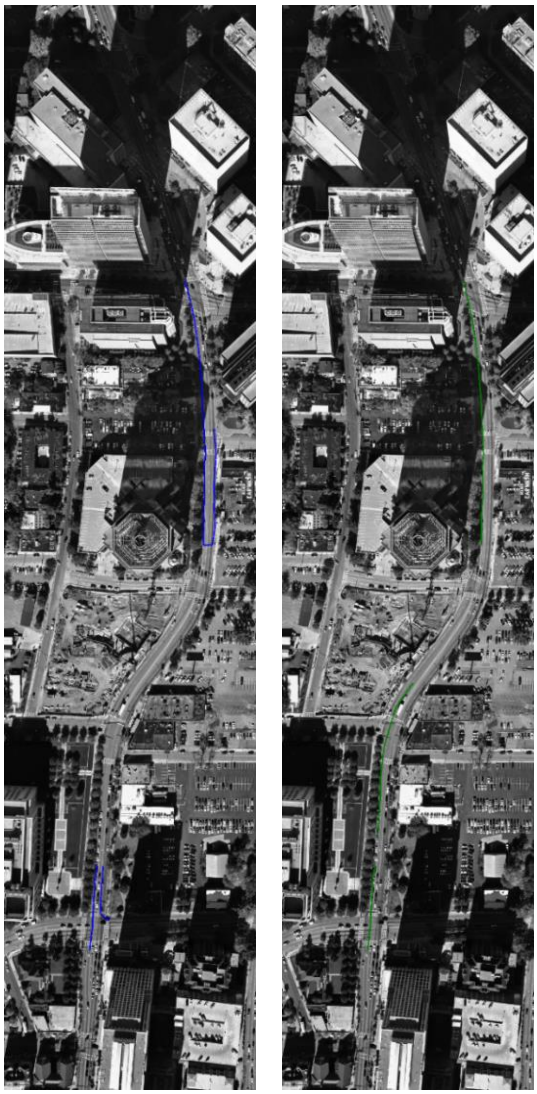
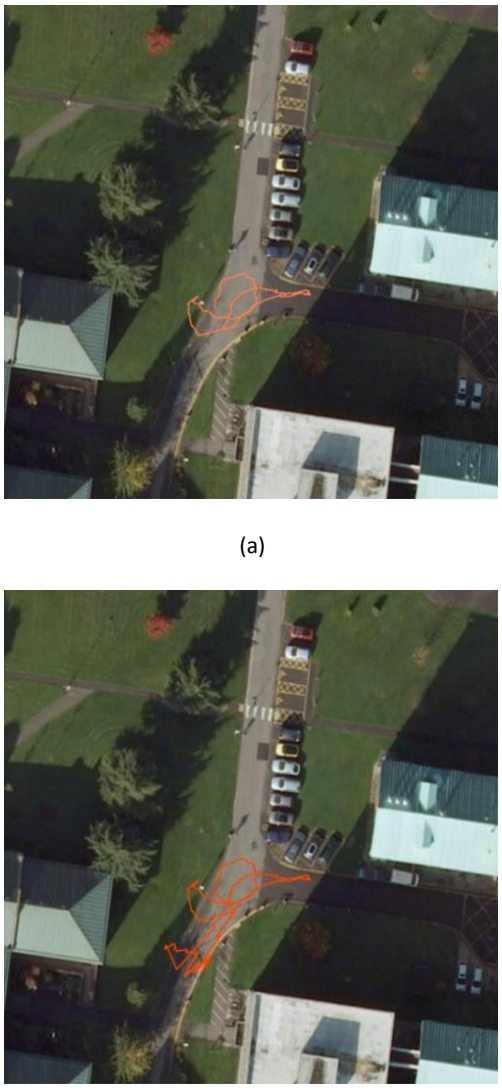


Fig. 10: Graph comparing the results of the proposed model with the other two models.



(a) (b)

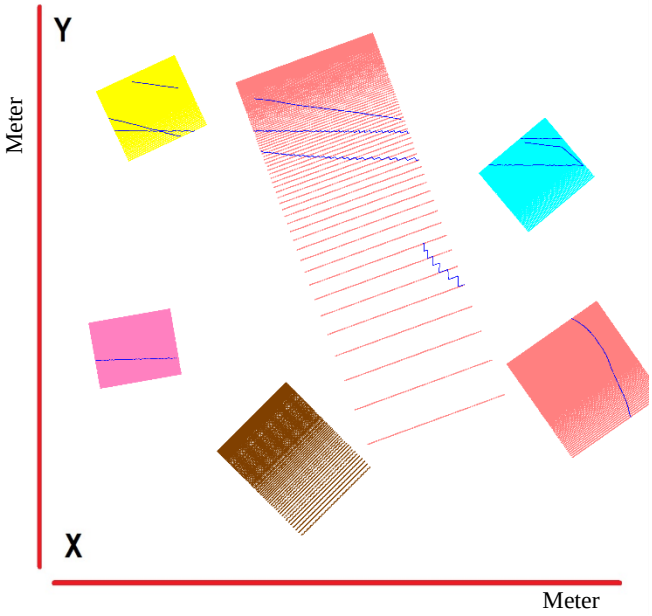
Fig. 11: The persistent tracking results of the NGSIM for one object: a) Extracted result; and b) Ground truth result.



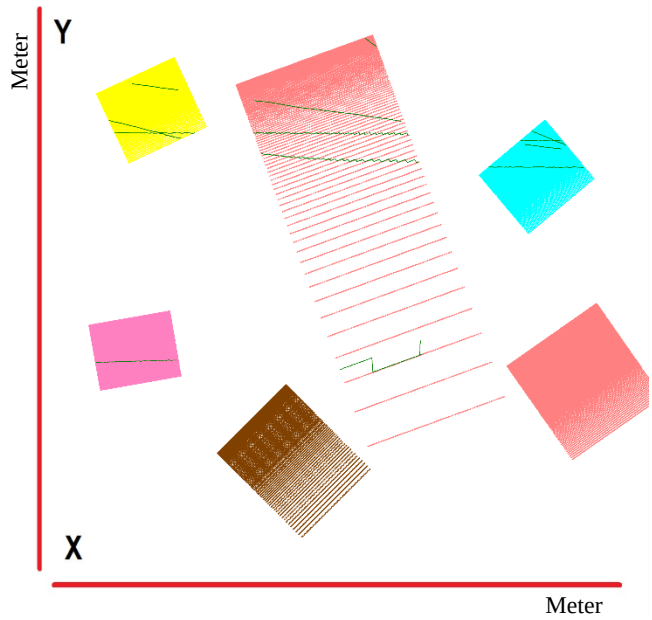
(a)

(b)

Fig. 12: The persistent tracking results of the PETS2009 for one object: a) Extracted result; and b) Ground truth result.



(a)



(b)

Fig. 13: The persistent tracking results of the synthesized data for one object: a) Extracted result; and b) Ground truth result.

Conclusion

In this paper, we proposed a variational multiphase model for associating the tracklets of the objects in the camera network and determining their persistent trace. We proposed a new representation of the multi-object tracking problem in a camera network which can solve this problem with less restricted assumption. So, this model is a more general model for multi-object tracking in camera network which doesn't need the serious prerequisite information of the wide area, the camera topology and the objects' models. We use the deep feature for appearance and motion representation of objects.

We have evaluated our proposed model by four complicated datasets using 10 well known and common metrics. These evaluations show the quality of the proposed model in solving complex problems using the minimum required initial knowledge.

Author Contributions

Ehsan Pazouki designed and simulated, carried out the data analysis and collected the data and Mohammad Rahmati interpreted the results and wrote the manuscript.

Acknowledgment

The authors gratefully thank the anonymous reviewers and the editor of JECEI for their useful comments and suggestions.

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

C_i	The *i*th Camera
C_{M_i}	Calibration Parameters of the *i*th Camera
P_τ	The τth Object
t_s	Start tracking time windows
t_e	End tracking time windows
T_C	*Tracklets* of all Cameras
T_{O_i}	*Tracklets* of the *i*th camera
R_P	Persistent trace of all objects
r_{P_x}	Persistent trace of the *x*th object
n	Count of moving objects
b	Count of persistent tracked of objects
ϕ	Multiphase Level Set representation of R_P
φ_τ	Level Set representation of r_{P_τ}
$\lvert \cdot \rvert$	The cardinal of the set

References

[1] A.K. Roy-Chowdhury, B. Song, Camera Networks: The Acquisition and Analysis of Videos over Wide Areas. Morgan & Claypool Publishers, 2012: 134.

[2] A. Yilmaz, O. Javed, M. Shah, "Object tracking: A survey," ACM Comput. Surv. (CSUR), 38(4):1-45, 2006.

[3] S. Challa, Fundamentals of object tracking. Cambridge, UK; New York: Cambridge University Press, 2011.

[4] J. Bins, L.L. Dihl, C. R. Jung, "Target tracking using multiple patches and weighted vector median filters," J. Math. Imaging Vision, 45(3): 293-307, 2013.

[5] Y. Sun, L. Bentabet, "A particle filtering and DSmT based approach for conflict resolving in case of target tracking with multiple cues," J. Math. Imaging Vision, 36(2): 159-167, 2010.

[6] G. Castanon, L. Finn, "Multi-target tracklet stitching through network flows," in Proc. IEEE Aerospace Conf., 1-7, 2011.

[7] J.-F. Aujol, "Calculus of variations in image processing," september 2008.

[8] A.G. Jagola, W. Yanfei, C. Yang, Computational Methods for Applied Inverse Problems. Berlin: De Gruyter, 2012.

[9] T.F. Chan, J.J.S.p. cm., Image processing and analysis: variational, PDE, wavelet, and stochastic methods. Siam: 400, 2005.

[10] G. Unal, A. Yezzi, "A variational approach to problems in calibration of multiple cameras," in Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR): I-172- I-178, 2004.

[11] N. Paragios, Y. Chen, O. Faugeras, Handbook of Mathematical Models in Computer Vision. Printed in the United States of America.: Springer, 2006.

[12] C. Liu, F. Dong, S. Zhu, D. Kong, K. Liu, "New variational formulations for level set evolution without reinitialization with applications to image segmentation," J. Math. Imaging Vision, 41(3): 194-209, 2011.

[13] O. Javed, Z. Rasheed, K. Shafique, M. Shah, "Tracking across multiple cameras with disjoint views," in Proc. Ninth IEEE International Conference on Computer Vision: 952-957, 2003.

[14] B. Song, A.K. Roy-Chowdhury, "Robust tracking in a camera network: A multi-objective optimization framework," IEEE IEEE J. Sel. Top. Signal Process., 2(4): 582-596, 2008.

[15] W. Hu, T. Tan, L. Wang, S. Maybank, "A survey on visual surveillance of object motion and behaviors," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., 34(3): 334-352, 2004.

[16] D. Makris, T. Ellis, J. Black, "Bridging the gaps between cameras," in Proc. of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition: II-205- II-210, 2004.

[17] S. C, K. Tieu, "Automated multi-camera planar tracking correspondence modeling," in Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition: I-259- I-266, 2003.

[18] R. Pless et al., "Persistence and tracking: Putting vehicles and trajectories in context," in Proc. 2009 IEEE Applied Imagery Pattern Recognition Workshop (AIPRW): 1-8, 2009.

[19] C. Zhu, "Multi-Camera People Detection and Tracking," Independent thesis Advanced level (degree of Master (Two Years)) Student thesis, 2019.

[20] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," Nature, 521(7553): 436-444, 2015.

[21] H.-M. Hsu, T.-W. Huang, G. Wang, J. Cai, Z. Lei, J. Hwang, "Multi-camera tracking of vehicles based on deep features Re-ID and trajectory-based camera link models," in CVPR Workshops, 2019.

[22] G. Wang, Y. Wang, H. Zhang, R. Gu, J.-N. Hwang, "Exploit the connectivity: Multi-Object Tracking with TrackletNet," ArXiv: 1811.07258, 2018.

[23] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," ArXiv:1512.03385, 2015.

[24] M.P. Ghaemmaghami, "Tracking of humans in video stream Using LSTM recurrent neural network," Master in Machine Learning, School of Computer Science And Communication, KTH Royal Institute of Technology School of Computer Science And Communication, 2019.

[25] D. Gordon, A. Farhadi, D. Fox, "Re3 : Real-time recurrent regression networks for object tracking," ArXiv: 1705.06368, 2017.

[26] P. Voigtlaender, J. Luiten, P. Torr, B. Leibe, "Siam R-CNN: visual tracking by re-detection," in Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 6577-6587, 2020.

[27] C. Ma et al., "Trajectory factory: Tracklet cleaving and re-connection by deep siamese bi-gru for multiple object tracking," ArXiv:1804.04555 [cs], 2018.

[28] X. Zhang, X. Wang, C. Gu, "Online multi-object tracking with pedestrian re-identification and occlusion processing," Visual Comput., 2020.

[29] N. Hussain et al., "A deep neural network and classical features based scheme for objects recognition: an application for machine inspection," Multimed. Tool. Appl., 2020: 1-23, 2020.

[30] M.A. Khan et al., "Human action recognition using fusion of multiview and deep features: an application to video surveillance," Multimed. Tool. Appl., 2020: 1-27, 2020.

[31] M. Rashid et al., "A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection," Sustainability, 12(12): 5037, 2020.

[32] M. Rashid, M.A. Khan, M. Sharif, M. Raza, M.M. Sarfraz, F. Afza, "Object detection and classification: a joint selection and fusion strategy of deep convolutional neural network and SIFT point features," Multimed. Tool. Appl., 78(12): 15751-15777, 2019.

[33] E. Pazouki, M. Rahmati, "Variational method for wide area surveillance," J. Ambient Intell. Smart Environ., 8: 189-203, 2016.

[34] E. Pazouki, M. Rahmati, "Multiphase vs. single-phase variational level set approach for video data association," Intell. Data Anal., 20: 679-699, 2016.

[35] R. Mohammdi Farsani, E. Pazouki, "A transformer self-attention model for time series forecasting," J. Electr. Comput. Eng. Innovations (JECEI), 9(1): 1-10, 2021.

[36] B. Dacorogna, Introduction to the Calculus of Variation. World Scientific Publishing Company, 2004: 240.

[37] T.F. Chan, L.A. Vese, "Active contours without edges," IEEE Trans. Image Process., 10(2): 266 – 277, 2001.

[38] "CAVIAR 2003 and 2004", accessed 23 February 2021.

[39] B. Song, R.J. Sethi, "Robust wide area tracking in single and multiple views," Rev. Lit. arts Am., 2011: 1-18, 2011.

[40] "ngsim peachtree street." accessed 23 February 2021.

[41] "Eleventh ieee international workshop PETS." accessed 23 February 2021.

[42] "Image Processing & Pattern Recognition Laboratory." accessed 23 February 2021.

[43] S. Inria, "Internal Technical note Metrics Definition version 2.0 – Approved," Inria, IN_ETI_1_004, 2006.

[44] Y. Li, C. Huang, R. Nevatia, "Learning to associate: HybridBoosted multi-target tracker for crowded scene," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 2953-2960, 2009.

Biographies

Ehsan Pazouki is a Professor in the school of Computer Engineering at Shahid Rajaee Teacher Training University where he has been a faculty member since 2016. Ehsan completed his Ph.D. and M.S. at Amirkabir University. His research interests lie in the area of wide area surveillance, ranging from theory to design to implementation. He has collaborated actively with researchers in several other disciplines of computer science, particularly Cognitive Science. Ehsan has experience in various industries related to his specialization for more than 10 years. For additional information see https://www.sru.ac.ir/en/school-of-computer/ehsan-pazouki/

Mohammad Rahmati received the M.Sc. degree in electrical engineering from the University of New Orleans, in 1987 and the Ph.D. degree in electrical and computer engineering from the University of Kentucky, Lexington, Kentucky, in 1994. He is currently an associate professor in the Computer Engineering Department, Amirkabir University of Technology (Tehran Polytechnic). His research interests include the fields of pattern recognition, image processing, bioinformatics, video processing, and data mining. He is the chair of the department and he is also a member of IEEE Signal Processing Society.

How to cite this paper:
E. Pazouki, M. Rahmati, "A variational level set approach to multiphase multi-object tracking in camera network base on deep features," J. Electr. Comput. Eng. Innovations, 9(2): 203-214, 2021.

DOI: 10.22061/JECEI.2021.7649.417

URL: https://jecei.sru.ac.ir/article_1542.html