International Journal of Engineering

Journal Homepage: www.ije.ir

Convolutional Gating Network for Object Tracking

A. Feizi*

Faculty of Electrical Engineering, Damghan University, Damghan, Semnan, Iran

PAPER INFO

Paper history:
Received 30 March 2019
Received in revised form 24 April 2019
Accepted 03 May 2019

Keywords:
Convolutional Neural Networks
Object Tracking
Convolutional Gating Network
Occlusion
Particle Filter

ABSTRACT

Object tracking through multiple cameras is a popular research topic in security and surveillance systems especially when human objects are the target. However, occlusion is one of the challenging problems for the tracking process. This paper proposes a multiple-camera-based cooperative tracking method to overcome the occlusion problem. The paper presents a new model for combining convolutional neural networks (CNNs), which allows the proposed method to learn the features with high discriminative power and geometrical independence. In the training phase, the CNNs are first pre-trained in each of the camera views, and a convolutional gating network (CGN) is simultaneously pre-trained to produce a weight for each CNN output. The CNNs are then transferred to the tracking task where the pre-trained parameters of the CNNs are re-trained by using the data from the tracking phase. The weights obtained from the CGN are used in order to fuse the features learnt by the CNNs and the resulting weighted combination of the features is employed to represent the objects. Finally, the particle filter is used in order to track objects. The experimental results showed the efficiency of the proposed method in this paper.

doi: *10.5829/ije.2019.32.07a.05*

1. INTRODUCTION

Research on visual surveillance systems is one of the most active topics in the field of computer vision. In such systems, single or multiple cameras and computers are installed to monitor behaviors of people, cars, or other objects, and consequently, these systems are employed in numerous applications including access control, person identification, congestion, crowd analysis, etc. Multi-view object tracking constitutes a major problem in areas under visual surveillance. The goal of object tracking is to follow a moving object over a sequence of images; however, the task is still challenging. The challenge is related to the variations in occlusion and appearance, which would cause surveillance systems unable to track objects with high accuracy. As the number of cameras in a surveillance system is increased, the system would achieve more complexity and higher accuracy in object tracking. Multi-view systems installed for object tracking are especially useful when occlusion occurs. Regardless of the number of the cameras, the basic idea behind all the object tracking systems is the same. A tracking system needs an object detection algorithm for when the object appears in the first frame. The tracked object needs to be represented by its appearance; this representation is usually called object feature. Once the object is detected and represented, its position is located in each of the new frames. The majority of the previous studies on object tracking have investigated surveillance systems with single cameras [1-6] using color information. However, in single-camera systems, the two factors of variable lighting conditions and occlusion would cause imperfect tracking results. On the other hand, when multiple-camera systems are employed, the area under surveillance is expanded, which lets information from multiple views be used to handle tracking issues such as occlusion. It is important to note that, in a multi-camera system, multiple cameras would cumulatively yield more discriminative information than any of the single cameras does. Exact representation of an object based on the extracted features is one of the most important steps in the process of tracking the object. An object can be more reliably tracked when the extracted features have high discriminative power, and the features are resistant to particular conditions such as lighting changes. Convolutional neural network (CNN) is one of the most

*Corresponding Author Email: *a.feizi@du.ac.ir* (A. Feizi)

Please cite this article as: A. Feizi, Convolutional Gating Network for Object Tracking, International Journal of Engineering (IJE), IJE TRANSACTIONS A: Basics Vol. 32, No. 7, (July 2019) 931-939

recently introduced methods for feature extraction. The features extracted through the CNN are distinguished by their high discriminative power.

This paper focuses on three issues which include 1) the collaboration of multiple cameras to track objects precisely and solve the problem of occlusion, 2) the extraction of features with high discriminative power and geometric independence through a combination of convolutional neural networks, and 3) using the extracted features as observations in order to estimate the location of objects more accurately during the process of tracking the objects via particle filters.

The remainder of this paper is organized as follows. Section 2 gives an overview of the literature on object tracking. Section 3 introduces the proposed method in this paper. Experimental results obtained in the paper are presented in section 4. Finally, section 5 concludes the paper.

1. 1. Related Work The purpose of the proposed method in this paper is to investigate the use of multiple cameras to track objects through the extraction of features by using CNN. In order to contextualize this purpose, the current section would first present a review of the available tracking methods based single or multiple cameras. The section would then go to discuss CNN and its applications in tracking systems. The majority of the previous studies have investigated the use of single cameras for object tracking [1-6].

Jin, and. Bhanu [1] employed a single camera is for multi-person tracking based on crowd simulation whereby the particle weights are modified in order to predict the position of a crowd by using a tracking-by-detection algorithm. Histogram Oriented Gradients (HOG)-based person detection is combined with a Gaussian per detection in order to address the problem of occlusion and track people by using single cameras [2]. The object detection and tracking are carried out by using a single camera based on graph-based flow optimization [3]. In this study, a template-based generative model is employed to perform detections by using a discretized ground plane. The tracking futsal players is performed based on the data from a single stationary camera [4]. In the study, adaptive background subtraction and blob analysis are carried out to detect the players. In addition, particle filters are used to track the players and predict their position. The study reported in [iterature [5] makes use of a Probability Hypothesis Density (PHD) filter to track the visual features which belong to a vehicle detected via a single camera. In another study [6], combination of a kernel-based histogram and a feature detector approach is proposed for the purpose of object tracking. The study employs two color spaces, i.e., RGB and HSI, and the kernel method to improve histogram accumulation.

A single camera does not suffice to achieve high accuracy in tracking objects under many conditions including object occlusion in the video frames, unstable lighting conditions, and complex background. Due to these limitations, it is difficult to handle occlusion in single cameras, and as a result, the process of object tracking is not accurately performed. In the recent years, tracking objects by using multiple cameras has attracted substantial attention as the use of multiple cameras would improve the process of object tracking in complex scenes. Furthermore, multiple cameras can be employed to solve the problem of occlusion. Consequently, object tracking via multiple cameras has been widely examined by the researchers in this field, who have adopted different approaches to this innovation. A method for tracking the position of multiple people in a scene by using overlapping cameras has been reported in literature [7]. In the study, a two-step approach that would simultaneously track people and estimate their position is proposed. Chen et al. [8] proposed the Switching Network Tracker (SNT), which would track multiple interacting targets (i.e., individuals and groups of people interacting with each other in natural scenes) by using both overlapping and non-overlapping cameras. The study reports the design of a Structural Support Vector Machine (SSVM) that integrates spatial and temporal relationships among the tracklets in order to detect the occurrence of group formation or splitting via a network of cameras. Lee et al. [9] have proposed a two-step Multiple Camera Tracking (MCT) method based on Single Camera Tracking (SCT) and a two-phase online feature learning Inter-Camera Tracking (ICT) method for the purpose of segmentation and local object detection. The study integrates three pose-invariant color features along a context-based couple feature with appearance cues to achieve this purpose. Yoon et al. [10] was introduced a Multiple Hypothesis Tracking (MHT) algorithm to track the targets across multiple cameras by maintaining the identities of the observations. By using a simple averaged color histogram as the appearance model, the method proposed in the study is able to track the targets both within and across the cameras [11]. The method proposed by Tesfaye et al. [12] adopts a unified three-layer hierarchical approach to solve tracking problems in multiple non-overlapping cameras. The method respectively employs the first two layers and the third layer of the approach to simultaneously solve the issues related to within-camera and across-camera tracking by merging the tracks of the same person in all the cameras.

CNN is used in numerous applications including face detection, medical image analysis, object recognition, image classification, and speech recognition. This technique, which is able to extract features from raw input images, is usually used to learn appearance models

from multi-cameras. There are a number of studies which have employed CNN for the purpose of object tracking.

Wang and Yeung [13] have proposed a tracking method based on denoising autoencoder is proposed. The method first learns image features and the learned features are then transferred to the online tracking task. The method introduced by Chen et al. [14] suggests a deep learning architecture which is able to dynamically learn the most discriminative features through CNN. Hong et al. [15] also proposed a pre-trained model, which makes use of CNN to perform the learning of discriminative detection. The method proposed by Zhang et al. [16] performs online object tracking based on deep learning in order to qualify sampling region localization,. To serve this purpose, the method integrates intra-frame appearance correlations and inter-frame motion saliency into a compositional energy optimization process. The ultimate output of the target localization along with the optional strategies demonstrates that the proposed method is effective in guiding a qualified sampling process for tracking by learning via CNN. The method introduced by Wang et al. [17] introduces a novel visual tracking algorithm in which three weak CNN trackers from the hierarchical convolution layers are combined with a stronger one. The method proposes a semiadaptive weighted convolutional features (SACF) algorithm to examine the performance of each weak tracker [18]. The method proposed in literature [19] integrates the extreme learning machine (ELM) autoencoder architecture into a CNN model and designs an updating scheme for the purpose of model training in order to overcome the tracking drift problem. The tracking task is decomposed into translation estimation and scale estimation is reported in literature [20]. In the process of translation estimation, multiple adaptive correlation filters are employed along with the CNN features in order to estimate the target location more accurately.

2. OUR CONTRIBUTION

In this section, the overall scheme of the proposed method in this paper is first explained, and the theoretical details of the method are then presented.

2. 1. Overal Scheme In designing the proposed method, the first step is to train and estimate the CNN and CGN parameters. To do so, the CNN is first pre-trained in each camera view. During the process of pre-training the CNN, each network is initially considered as a classifier. In the input frames, the objects are manually labeled and then used to train the networks. It is necessary to mention that the parameter configuration of the CNN is the same as the image classification model of the VGG-19-layer [21]. In the current paper, a CGN is simultaneously trained along with the CNN. Based on the confidence degrees of the cameras, the CGN is used to produce weights in order to fuse the features extracted from the cameras. After the pre-training phase, the CNNs are transferred to the tracking phase during which the pre-trained VGG-19 model is fine-tuned by training the data obtained from the latter phase. The schematic representation of the proposed method in the training phase is illustrated in Figure 1. Once the CNN parameters are trained and estimated in each camera view, the networks are exited from the classifier system by deleting the last layer, which are then considered as a feature extraction system for object tracking.

The particle filter is a probabilistic filter that estimates the state of an object in the current frame based on its state in the previous frames, and also based on the measured observations in the current frame. In this paper, the objects extracted through the CNNs are used as the observations in order to improve the state of the objects in the particle filters. The proposed method in this paper has the following properties:

1. Once an instance of occlusion occurs in a camera view, the features with the higher degrees of confidence will improve the occlusion problem.
2. The features extracted in each camera view via the CNN have high discriminative power. These features also have geometric independence, and therefore, they can be used in multiple camera systems.
3. During the process of tracking, CGNs let the features of the more reliable cameras claim a more important role in representing the target. This provides the opportunity to represent and track the target more accurately.

2. 2. CNN In the recent years, CNN-based feature extraction has made tracking alghorithms achieve significant improvements [22]. LeCun et al. [23] proposed CNN, which is a model of neural networks, the aim of which is to learn features from image pixels directly. In CNN there are three main types of layers; i.e., convolution layers, max-pooling (or sub-sampling) layers, and a fully connected layer. The convolution and max-pooling layers are considered as 2-D layers while the fully connected layer is considered as a 1-D layer. Each of the convolution layers is usually followed by the activation units that rescale the results of the convolution in a nonlinear manner. The output of a convolution layer is called a feature map. Through different methods such as average-pooling or max-pooling, the pooling layers reduce the dimensionality of the feature maps created by the convolutional layers. Finally, the fully connected layer is used to extract high level features from the data in the CNN, the user specifies the network architecture by defining the number of the layers, their kinds, and the type of the activation units. In this paper, the CNN architecture of the VGG-19 network [21] is employed for the purpose of feature extraction and pre-training using the large scale dataset ImageNet. In the VGG-19 network there are 19 learnable layers (see Figure 2).

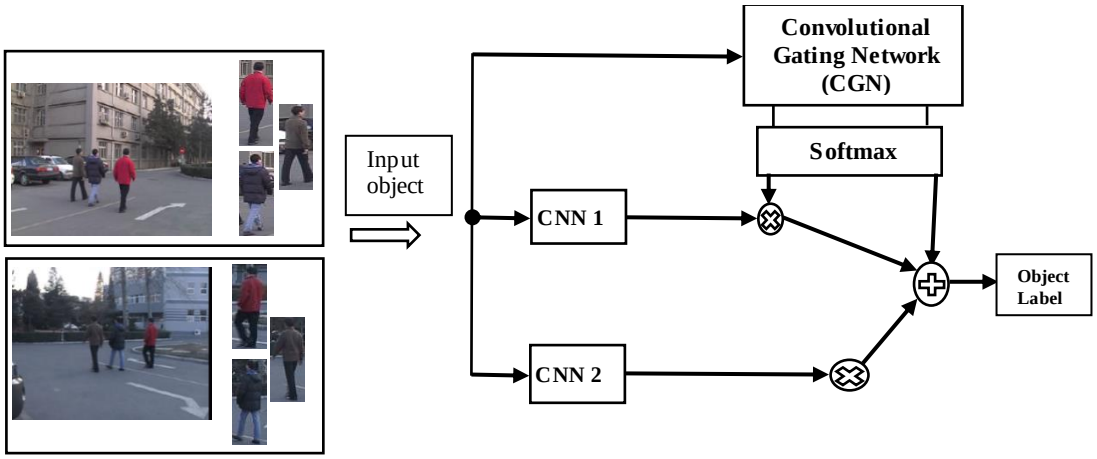

Figure 1. The flowchart of the proposed algorithm

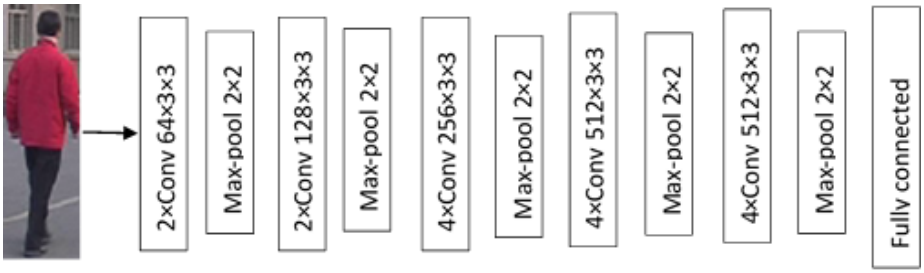

Figure 2. The CNN architecture of the VGG-Net-19 network

2. 3. CNN and CGN Training This paper proposes a method that combines CNNs for the purpose of extracting the features of an object. Let K be the camera view. For each camera view, one CNN is set, and thus, there are K CNNs and one CGN, which share the same input. With training the CGN input-adaptive weights are produced and then are used for fusing the outputs of the CNNs. Note that the CNNs and CGN are simultaneously trained. Consider the training set of M object images. Following literature [24], the error function of the CNNs for the m-th input object image is defined as follows:

$$E^m = -\ln\left(\sum_{k=1}^{K} p_k^m \exp\left(-\tfrac{1}{2}\left\|y^m - O_k^m\right\|^2\right)\right) \quad (1)$$

where O_k^m represents the output vector of the k-th CNN and y^m represents the desired output vector. For the k-th CNN, the effective error signal for all the training samples can be written as follows:

$$e_k = \sum_{m=1}^{M} \frac{\partial E^m}{\partial O_k^m} = \sum_{m=1}^{M} q_k^m (y^m - O_k^m) \quad (2)$$

$$q_k^m = \frac{p_k^m \exp\left(-\tfrac{1}{2}\left\|y^m - O_k^m\right\|^2\right)}{\sum_{j=1}^{K} p_j^m \exp\left(-\tfrac{1}{2}\left\|y^m - O_j^m\right\|^2\right)} \quad (3)$$

In this paper, the total error function for the convolutional gating network is defined as follows:

$$E_{CGN} = \frac{1}{2}\sum_{m=1}^{M}\left\|q^m - p^m\right\| \quad (4)$$

where $q^m = \left[q_1^m, q_2^m, ..., q_K^m\right]$ represents a vector of the posterior-probability estimation produced by the CNNs for the m-th input object image and $p^m = \left[p_1^m, p_2^m, ..., p_K^m\right]$ represents the CGN output vector. Once the error signal for the CNNs and the error function for the CGN are defined, the batch training algorithm proposed in literature [25] (RPROP) is used to simultaneously determine the free parameters of the CNNs and CGN. RPROP is a fast back propagation variant similar in spirit to Quick prop. It is about as fast as Quick prop but requires less adjustment of the parameters to be stable. The parameters used were not determined by a trial-and-error search, but are just educated guesses instead. RPROP requires epoch learning, i.e., the weights are up dated only once per epoch. While epoch updates are is not desirable for very large training sets, it is a good method for small and medium training sets. Riedmiller and Braun [25] claimed that this algorithm is faster converged in compare to the other algorithms, and it also works well with gradients of small magnitude.

2. 4. Tracking Tracking can be simply defined as the estimation of the path of an object moving over an image sequence of a video. In the field of behavior recognition, the main problem is to establish motion correspondences across consecutive video frames. The estimated path of the movement of an object during the tracking process is called as the trajectory of the object.

In this section, the CNN-based appearance model is integrated into the particle filtering framework in order to track objects. In the particle filter, a recursive Bayesian filter via the Monte Carlo sampling method is implemented [26]. In the particle filter, the posterior density of the state of the object is represented by a set of random particles along with their associated weights. The two main components in the particle filter include a dynamic model, which is employed to generate the candidate samples based on the previous particles, and an observation model, which is employed to calculate the similarity between the candidate samples and the object appearance model.

Given that all the measures (observations) of the object up to the time t are obtained, $z_{1:t}=[z_1,...,z_t]$, the posterior density $P(x_t|z_{1:t})$ can be reformulated as follows:

$$p(x_t|z_{1:t})=p(z_t|x_t)\int p(x_t|x_{t-1})p(x_{t-1}|z_{1:t-1})dx_{t-1} \quad (5)$$

where $p(x_t|x_{t-1})$ represents the dynamic model, and $p(z_t|x_t)$ represents observation model.

In the particle filter, the posterior density $P(x_t|z_{1:t})$ is represented by a set of samples (particles) as follows:

$$\left\{s_t^{(n)}:n=1,...,N\right\} \quad (6)$$

with their associated weights (sampling probability) as follows:

$$\left\{\pi_t^{(n)}:n=1,...,N\right\} \quad (7)$$

Finally, the optimal state of the object at the time t is determined by the maximum posterior estimation as follows:

$$\begin{aligned} x_t &= \arg\max_{x_t} p(x_t|z_{1:t})=x_t^n \\ &= \arg\max_{x_t^n} \pi_t^n \end{aligned} \quad (8)$$

3. TESTS AND RESULTS

In this paper, the proposed method is evaluated on three standard video datasets. Comparisons are also made between the proposed method and a number of other methods for tracking objects. In this section, the datasets are first introduced, and the experimental results are then presented.

3. 1. Implementation The proposed method in this paper is implemented through MATLAB. In addition, all the experiments are performed by using a workstation with 3.2 GHz Intel i5 processor and 8 GB RAM. In this paper, the same frame detection obtained by the DPM algorithm reported in literature [27] was used as the input to all the tracking methods under comparison and the number of the particle filters is set to 1000. The target objects are extracted and normalized to 32×32 patches in all the images.

3. 2. Datasets Three public datasets are employed in this paper as follows:

PETS 2009 dataset: This dataset is one of the most commonly used datasets for evaluating tracking tasks. In this dataset, the sequence S2/L1 is specially designed for multi-view-based tasks whereby the movements of 10 pedestrians are tracked. The frame size and rate of the targets are 720 × 576 pixels and 7 fps, respectively. In this paper, view 001 and view 007 are used for object tracking (2 cameras) [28].

PETS 2001: The PETS 2001 dataset contains four separate sets of sequences. Of the four sets, Dataset 1, which contains multi-view (2 cameras) video sequences with a resolution of 720×576, is sampled for the purpose of the present study [29].

CAVIAR dataset: In this dataset, six basic scenarios are acted out by the CAVIAR team members. A subset of this dataset contains videos recorded by two cameras from two different views with a resolution of 384×288 pixels and a frame rate of 25 fps. In the subset, the first camera shows a view of the corridor while the second camera shows a frontal view of the scenario [30].

The qualitative results for the three dataset under comparison in this paper are shown in Figure 3. In the figure, each row shows the results of the tracking in the dataset. The first and second columns in the figure present the results obtained for the two sample frames from the first view whereas the third and fourth columns present the obtained results for the two sample frames from the second view.

3. 3. Evaluation In this paper, the proposed method is evaluated via the task of object matching precision. Correct tracking (i.e., the number of the frames that correct object tracking is performed), false tracking (i.e., the number of the frames in which the object is considered as a new object) and missing tracking (i.e., the number of the frames that missed tracking is performed), are employed as the target measures in order to determine the extent of object matching precision.

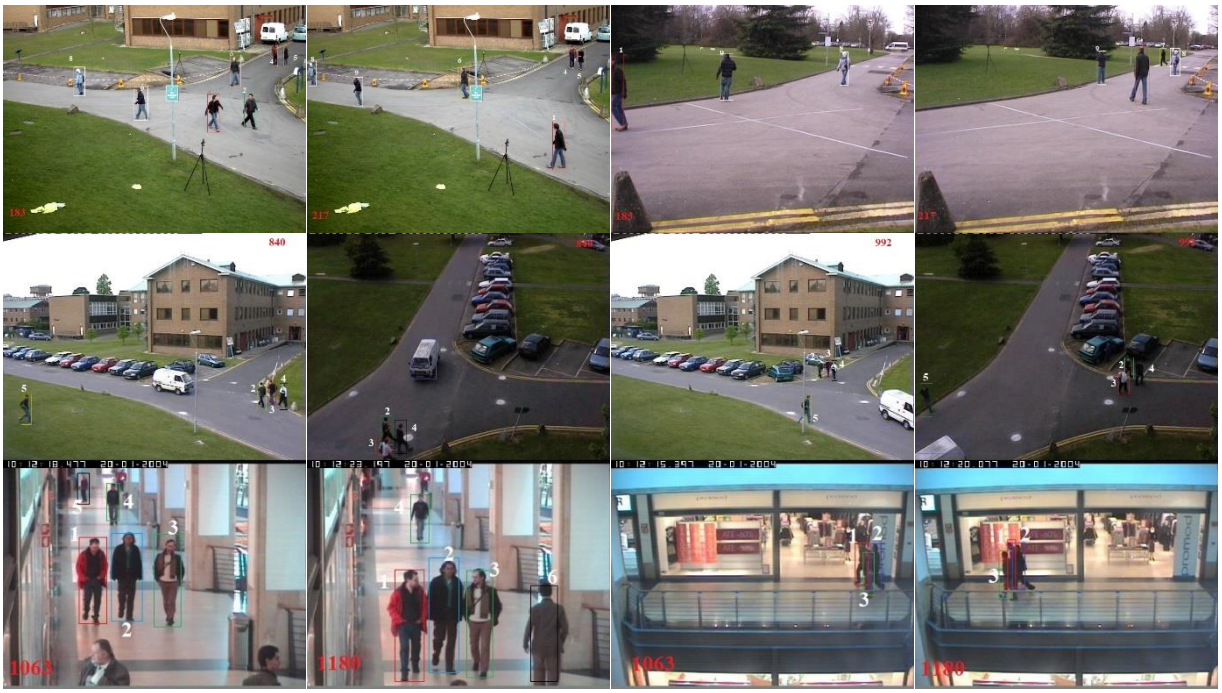

Figure 3. Tracking samples in the PETS 2009 (first row), PETS 2001(second row) and CAVIAR (third row) datasets

The proposed method in this paper is also evaluated based on the distance between the real object and the tracked object. In this latter evaluation, the number of the frames with successful tracking is counted. To do so, multiple object tracking precision (MOTP), the metrics proposed by Stiefelhagen et al. [31], and multiple object tracking accuracy (MOTA), are adopted for the purpose of evaluation.

Following literature [32], the greedy strategy within a distance threshold is used to evaluate the performance of the proposed method. The strategy tries to match the locations of the tracked objects and the ground truth. The MOTA scores, which represent the performance of the tracker within the distance thresholds varying from 0 to 2 m, is also plotted.

Occlusion is one of the critical problems for object tracking in cluttered scenes. Once occlusion occurs, the features extracted from the tracked object are disrupted, which in return makes the tracking process more complicated. This is especially the case with respect to single-camera systems since no additional information is available to improve the results of the tracking process. Consequently, the proposed method in this paper makes use of multiple cameras in order to address the problem of occlusion. To evaluate the proposed method and compare it with the rival methods on the problem of occlusion, the PETS 2009 dataset that represent challenges examples of frequent target occlusions is employed.

In Table 1 and Figure 4, the comparative results for the proposed method in this paper vis-à-vis several state-of-the-art multi-camera tracking methods with respect to addressing the problem of occlusion on the PETS 2009 dataset are presented.

A distance threshold of 1 m is set to calculate the results for the target methods as shown in Table 1. Table 1 shows that the results of the tracking process are improved when the proposed method are implemented to the dataset. It should be noted that the performance of the proposed method in tracking human objects in crowded scenes is considerably better in compare to other target methods, which is ascribed to the accuracy of the features extracted through the proposed method.

TABLE 1. Performance of the tracking methods in the PETS 2009 S2.L1 dataset

Method	Correct Tracking	Missing Tracking	False Tracking	MOTA	MOTP
[33]	3993	95	84	89.05	78.10
[34]	4002	78	92	91.57	80.30
[35]	3985	100	87	84.90	67.95
[36]	4085	45	42	95.44	80.80
[37]	4010	75	87	90.1	81.04
Proposed Method	4100	40	32	97.2	82.4

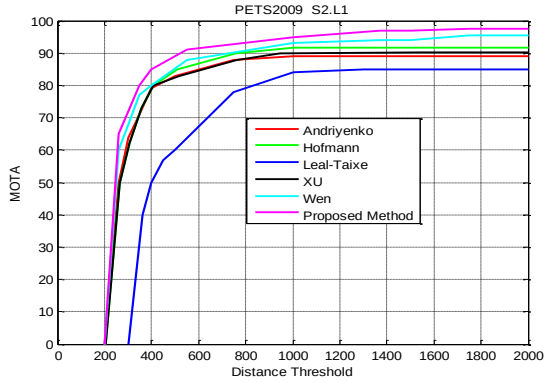

Figure 4. Plots of the MOTA scores within different distance thresholds in the PETS 2009 S2.L1 dataset

To evaluate performance, the proposed approach is compared with several methods. Andriyenko [33] proposed to track objects in discrete space and used splines to model trajectories in continuous space. The proposed method by Hofmann et al. [34], formulated a unified hierarchical multi object tracking architecture. Tracking of multiple objects simultaneously is cast as a MAP problem which is solved via a three-stage framework. Explicit occlusion reasoning is also considered. Leal-Taixé et al. [35] presents a novel platform for evaluating multi-target tracking approaches. The presented method by Wen et al. [36], proposed a new multi camera multi-target tracking method based on a space-time-view hyper-graph that encoded higher-order constraints (i.e., beyond pairwise relations) on 3D geometry, appearance, motion continuity, and trajectory smoothness among 2D tracklets within and across different camera views. The proposed method by Xu et al. [37], formulated multi-view multi-object tracking as a structure optimization problem described by a hierarchical composition model. The objective is to discover composition gradients of each object in the hierarchical graph.

Table 2 and Figure 5 show the comparative results for the target methods with respect to the PETS 2001 dataset. The comparative results for a distance threshold of 1 m are obtained. In this dataset, the value of the correct fraction for the proposed method in this paper is respectively 1, 16, and 10% higher than those for the tracking methods [34, 36, 37].

The CAVIAR dataset is another challenging dataset for object tracking with overlapping Field of Views (FOVs), which includes instances of object occlusion and low frame rate. The results for the proposed method against those for the rival methods in this dataset with distance threshold 1 m are presented in Table 3 and Figure 6. The obtained results showed that the average MOTA for the proposed method in this paper is respectively 6, 32, and 23% higher than those for the tracking methods in [32, 34, 35].

TABLE 2. Performance of the tracking methods in the PETS 2001 dataset

Method	Correct Tracking	Missing Tracking	False Tracking	MOTA	MOTP
[33]	2219	125	168	73.1	68.6
[34]	2240	111	161	76.5	71.5
[35]	2154	178	180	69.4	60.5
[36]	2451	32	29	90.31	78.5
[37]	2384	62	66	82.4	67.6
Proposed Method	2472	21	19	92.2	81.53

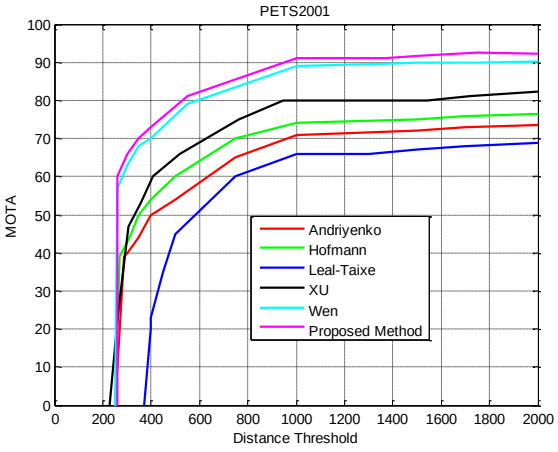

Figure 5. Plots of the MOTA scores within different distance thresholds in the PETS 2001 dataset

TABLE 3. Performance of the tracking methods in the CAVIAR dataset

Method	Correct Tracking	Missing Tracking	False Tracking	MOTA	MOTP
[33]	1940	40	33	65.61	58.57
[34]	1938	43	32	62.5	55.71
[35]	1907	50	56	57.26	48.46
[36]	1971	23	19	82.92	75.48
[37]	1953	32	28	71.4	63.5
Proposed Method	1985	17	11	88.6	80.45

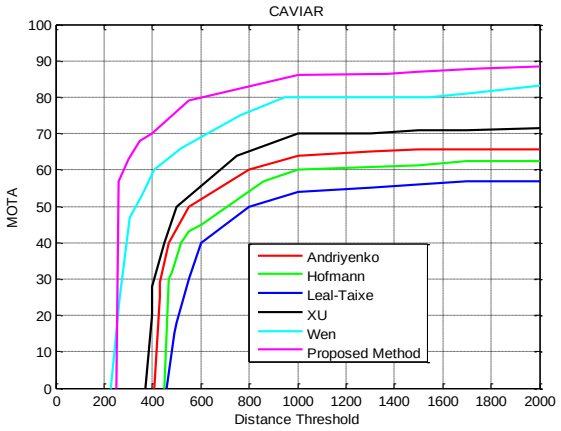

Figure 6. Plots of the MOTA scores within different distance thresholds in the CAVIAR dataset

4. CONCLUSION

In this paper, a tracking method based on multiple cooperative cameras is proposed in order to overcome the occlusion problem. For this purpose, a new model for combining CNNs, which allows the proposed method to learn image features more efficiently is suggested. The results indicate that the extracted features have high discriminative power and geometrical independence. In the training phase CNNs and a CGN are first pre-trained in each camera view simultaneously. The CNNs are then transferred to the tracking task where the pre-trained parameters of the CNNs are re-trained through the data from the tracking phase. The weights obtained from the CGN are used in order to fuse the features learnt by the CNNs, and the resulting weighted combination of the features is used to represent objects. Finally, the particle filter is used to track the objects. To conclude, the proposed method in this paper is capable of tracking objects as effectively as possible.

5. REFERENCES

1. Jin, Z. and Bhanu, B., "Single camera multi-person tracking based on crowd simulation", in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE. V, (2012), 3660-3663.
2. Andriyenko, A., Roth, S. and Schindler, K., "An analytical formulation of global occlusion reasoning for multi-target tracking", in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE., (2011), 1839-1846.
3. Wu, Z., Thangali, A., Sclaroff, S. and Betke, M., "Coupling detection and data association for multiple object tracking", in 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE., (2012), 1948-1955.
4. de Pádua, P.H., Pádua, F.L., Sousa, M.T. and Pereira, M.dA., "Particle filter-based predictive tracking of futsal players from a single stationary camera", in 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, IEEE., (2015), 134-141.
5. García, F., Prioletti, A., Cerri, P. and Broggi, A., "Phd filter for vehicle tracking based on a monocular camera", ***Expert Systems with Applications***, Vol. 91, (2018), 472-479.
6. Zulkifley, M.A., "Robust single object tracker based on kernelled patch of a fixed rgb camera", ***Optik-International Journal for Light and Electron Optics***, Vol. 127, No. 3, (2016), 1100-1110.
7. Liem, M.C. and Gavrila, D.M., "Joint multi-person detection and tracking from overlapping cameras", ***Computer Vision and Image Understanding***, Vol. 128, (2014), 36-50.
8. Chen, X., Qin, Z., An, L. and Bhanu, B., "An online learned elementary grouping model for multi-target tracking", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., (2014), 1242-1249.
9. Lee, Y.-G., Tang, Z. and Hwang, J.-N., "Online-learning-based human tracking across non-overlapping cameras", ***IEEE Transactions on Circuits and Systems for Video Technology***, Vol. 28, No. 10, (2017), 2870-2883.
10. Yoon, K., Song, Y.-m. and Jeon, M., "Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views", ***IET Image Processing***, Vol. 12, No. 7, (2018), 1175-1184.
11. Lin, D.-T. and Huang, K.-Y., "Collaborative pedestrian tracking and data fusion with multiple cameras", ***IEEE Transactions on Information Forensics and Security***, Vol. 6, No. 4, (2011), 1432-1444.
12. Tesfaye, Y.T., Zemene, E., Prati, A., Pelillo, M. and Shah, M., "Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets", arXiv preprint arXiv:1706.06196, (2017), 1-15.
13. Wang, N. and Yeung, D.-Y., "Learning a deep compact image representation for visual tracking", in Advances in neural information processing systems., (2013), 809-817.
14. Chen, Y., Yang, X., Zhong, B., Pan, S., Chen, D. and Zhang, H., "Cnntracker: Online discriminative object tracking via deep convolutional neural network", ***Applied Soft Computing***, Vol. 38, (2016), 1088-1098.
15. Hong, S., You, T., Kwak, S. and Han, B., "Online tracking by learning discriminative saliency map with convolutional neural network", in International conference on machine learning., (2015), 597-606.
16. Zhang, P., Zhuo, T., Huang, W., Chen, K. and Kankanhalli, M., "Online object tracking based on cnn with spatial-temporal saliency guided sampling", ***Neurocomputing***, Vol. 257, (2017), 115-127.
17. Wang, H., Zhang, S., Ge, H., Chen, G. and Du, Y., "Robust visual tracking via semiadaptive weighted convolutional features", ***IEEE Signal Processing Letters***, Vol. 25, No. 5, (2018), 670-674.
18. Qian, X., Han, L., Wang, Y. and Ding, M., "Deep learning assisted robust visual tracking with adaptive particle filtering", ***Signal Processing: Image Communication***, Vol. 60, (2018), 183-192.
19. Sun, R., Wang, X. and Yan, X., "Robust visual tracking based on extreme learning machine with multiple kernels features fusion", in 2017 3rd IEEE International Conference on Computer and Communications (ICCC), IEEE., (2017), 2029-2033.
20. Hao, Z., Liu, G. and Zhang, H., "Correlation filter-based visual tracking via adaptive weighted cnn features fusion", ***IET Image Processing***, Vol. 12, No. 8, (2018), 1423-1431.
21. Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, (2014).

22. Wu, G., Lu, W., Gao, G., Zhao, C. and Liu, J., "Regional deep learning model for visual tracking", ***Neurocomputing***, Vol. 175, (2016), 310-323.
23. LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., "Gradient-based learning applied to document recognition", ***Proceedings of the IEEE***, Vol. 86, No. 11, (1998), 2278-2324.
24. Jacobs, R.A., Jordan, M.I., Nowlan, S.J. and Hinton, G.E., "Adaptive mixtures of local experts", ***Neural computation***, Vol. 3, No. 1, (1991), 79-87.
25. Riedmiller, M. and Braun, H., "A direct adaptive method for faster backpropagation learning: The rprop algorithm", in Proceedings of the IEEE international conference on neural networks, San Francisco. Vol. 1993, (1993), 586-591.
26. Arulampalam, M.S., Maskell, S., Gordon, N. and Clapp, T., "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking", ***IEEE Transactions on signal processing***, Vol. 50, No. 2, (2002), 174-188.
27. Felzenszwalb, P.F., McAllester, D.A. and Ramanan, D., "A discriminatively trained, multiscale, deformable part model", in Cvpr. Vol. 2, (2008).
28. "Pets2009.Available: Http://www.Cvg.Reading.Ac.Uk/pets2009/a.Html."
29. *Pets2001*.Available: http://www.cvg.reading.ac.uk/PETS2001/pets2001-dataset.html.
30. "Caviar.Available: Http://groups.Inf.Ed.Ac.Uk/vision/caviar/caviardata1/."
31. Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D. and Soundararajan, P., "The clear 2006 evaluation", in International evaluation workshop on classification of events, activities and relationships, Springer., (2006), 1-44.
32. Andriyenko, A., Schindler, K. and Roth, S., "Discrete-continuous optimization for multi-target tracking", in 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE., (2012), 1926-1933.
33. Andriyenko, A. and Schindler, K., "Multi-target tracking by continuous energy minimization", in CVPR 2011, IEEE., (2011), 1265-1272.
34. Hofmann, M., Wolf, D. and Rigoll, G., "Hypergraphs for joint multi-view reconstruction and multi-object tracking", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., (2013), 3650-3657.
35. Leal-Taixé, L., Milan, A., Reid, I., Roth, S. and Schindler, K., "Motchallenge 2015: Towards a benchmark for multi-target tracking", arXiv preprint arXiv:1504.01942, (2015).
36. Wen, L., Lei, Z., Chang, M.-C., Qi, H. and Lyu, S., "Multi-camera multi-target tracking with space-time-view hyper-graph", ***International Journal of Computer Vision***, Vol. 122, No. 2, (2017), 313-333.
37. Xu, Y., Liu, X., Liu, Y. and Zhu, S.-C., "Multi-view people tracking via hierarchical trajectory composition", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., (2016), 4256-4265.

Convolutional Gating Network for Object Tracking

A. Feizi

Faculty of Electrical Engineering, Damghan University, Damghan, Semnan, Iran

P A P E R I N F O

Paper history:
Received 30 March 2019
Received in revised form 24 April 2019
Accepted 03 May 2019

Keywords:
Convolutional Neural Networks
Object Tracking
Convolutional Gating Network
Occlusion
Particle Filter

چکیده

تعقیب هدف با استفاده از چند دوربین یک موضوع تحقیقاتی مهم در سیستم های نظارتی و امنیتی است. با این وجود مساله انسداد هدف یک موضوع چالش برانگیز در فرایند تعقیب هدف است. در این مقاله یک سیستم تعقیب هدف بر اساس همکاری چند دوربین برای تعقیب هدف ارائه شده است. این مقاله یک روش جدید برای ترکیب شبکه های عصبی کانولوشنی ارائه می‌دهد که این روش منجر به استخراج ویژگی‌هایی با توان متمایز کنندگی بالا و خاصیت استقلال هندسی می گردد. در مرحله یادگیری، برای هر نمای دوربین یک شبکه عصبی کانولوشنی پیش‌آموزش داده می‌شود. همزمان یک شبکه CGN نیز برای تولید وزن‌ها پیش‌آموزش داده می‌شود. سپس این شبکه‌های پیش‌اموزش دیده به فرایند تعقیب هدف انتقال داده می‌شوند و ویژگی‌ها استخراج می‌شوند. در نهایت فیلتر ذرات برای تعقیب هدف با استفاده از این ویژگی‌های استخراج شده مورد استفاده قرار می گیرند. نتایج آزمایشی به دست آمده، کارآمد بودن روش پیشنهادی را نشان می‌دهد.

***doi**: 10.5829/ije.2019.32.07a.05*