https://doi.org/10.2352/ISSN.2470-1173.2020.16.AVM-108
© 2020, Society for Imaging Science and Technology

Multiple pedestrian tracking using Siamese random forests and shallow Convolutional Neural Networks

*Jimi Lee, Jaeyeal Nam, ByoungChul Ko**
Department of Computer Engineering, 1095, Dalgubeol-daero, Dalseo-gu, Daegu, Republic of Korea
jimi88lee@gmail.com, jynam@kmu.ac.kr, niceko@kmu.ac.kr (Corresponding author)

Abstract

In this study, we propose a new multi-pedestrian tracking (MPT) method that performs quickly and efficiently track pedestrians in real-time system. The proposed method considers combining shallow convolutional neural networks (CNN) with ensemble learning method, Siamese random forests (SRF). Unlike conventional methods, to promote robustness of ensemble method, feature transformation is applied which exploit shallow networks in appearances of still images to extract enrich features. We formulate the problem of MOT in a structured learning framework based on SRF. Each forest learns differences of random feature pairs, which are extracted from the former process to enhance robustness to easily happened circumstances in a moving vehicle. When it compares to the conventional tracking algorithms, the proposed approach, based on SRF, takes advantage of lightweight and efficiency. The proposed lightweight multiple pedestrian tracker was successfully applied to benchmark datasets and yielded a similar or better performance level as compared with state-of-the-art methods.

Introduction

Multi-Pedestrian tracking (MPT) is an important factor in advance driver assistant system (ADAS) because the crashing possibility is significantly reduced depending on whether the pedestrian tracking how fast. In ADAS, drivers can be alerted to potential risky pedestrians as early as possible to avoid a collision. Therefore, MPT is the basic step to reach the final level of autonomous vehicles (AV). However, MPT from a moving vehicle is very challenging such as dynamic backgrounds, camera movement, deformable appearances of pedestrians, and occlusions by other objects [1].

Therefore, this paper focuses on introducing fast and accurate pedestrian tracking technology that is one of the most important core technologies in AV. The main contributions and overall procedures of our study can be summarized as follows;

1) We introduce a new MPT algorithm based on Siamese random forests (SRF) tracker model using output features of a YOLOv3 in a moving vehicle in real time.

2) In this study, we propose SRF for similarity matching between detection and tracker as an alternative to the well-known Siamese neural networks.

3) The average performance of the proposed approach is higher that of other methods in terms of MOTP and computational speed.

Related Work

Traditional MPT methods follow a 'detection-by-tracking' approach. In this method, the pre-trained object detector detects pedestrians and predicts the pedestrian position in the next frame by means of mean-shift [12], kalman-filter [13], and particle-filter [1,14]. This method works well in a fixed camera environment, but performance is inferior because pedestrian movement is difficult to predict in a vehicle environment where the camera moves irregularly.

Most state-of-the-art MPT methods use a convolutional neural network (CNN) which have shown to produce good result without the need for manual feature extraction. Li et al. [2] proposed a target-specific CNN for object tracking; the CNN is re-trained incrementally during tracking with new examples obtained online. This approach tracks only one object and uses a CNN for online learning. Because online learning of CNNs incurs high-level computational complexity and the use of MPT requires an individual tracker model, this method is not feasible in real time. Online learning based on CNNs is not straightforward owing to the large network size and lack of training data. Therefore, several tracking approaches based on CNNs focus on data association using a network trained offline to determine whether two detections belong to the same trajectory [3][4].

Recently, tracking research using Siamese CNN, which applied person-identification research, has been actively conducted in the MPT field. Siamese CNN applies the same network to the detection object and tracker object and calculates similarity with the difference of the output feature value. Therefore, Siamese CNN does not need to maintain separate network structure and has the advantage of fast tracking.

Wang et al. [9] first trained a Siamese CNN on the auxiliary data, and then jointly learned an online Siamese CNN with temporally constrained metrics to construct appearance-based tracklet affinity models. For a reliable association between tracklets, a temporally constrained multi-task learning mechanism based on a novel loss function incorporation is proposed. Son et al. [4] proposed a quadruplet architecture of a deep neural network by modifying a Siamese CNN and a triplet network to learn the object association for the MPT. This method combines the appearance of detections with their sequence-specific motion-aware position for metric learning, and the entire network is trained end-to-end in a unified framework. Zhu et al. [10] proposed dual matching attention network by combining both spatial and temporal attention mechanisms. The spatial attention module using Siamese architecture to handle noisy detections and occlusions.

However, for real-time multiple pedestrian tracking, Siamese network has some problems when fully connected network (FCN) is used for similarity measure. First, the two shared base networks must run in parallel at the same time. Second, the training data for base network are not yet sufficient. Three, the deep neural network requires a large amount of computational time.

The purpose of this paper is to propose a new fast and lightweight MPT system that is essential for AV safety. The proposed SRF is capable of real-time MPT compared to the existing CNN-based MPT system, and has the advantages of ensemble technique for

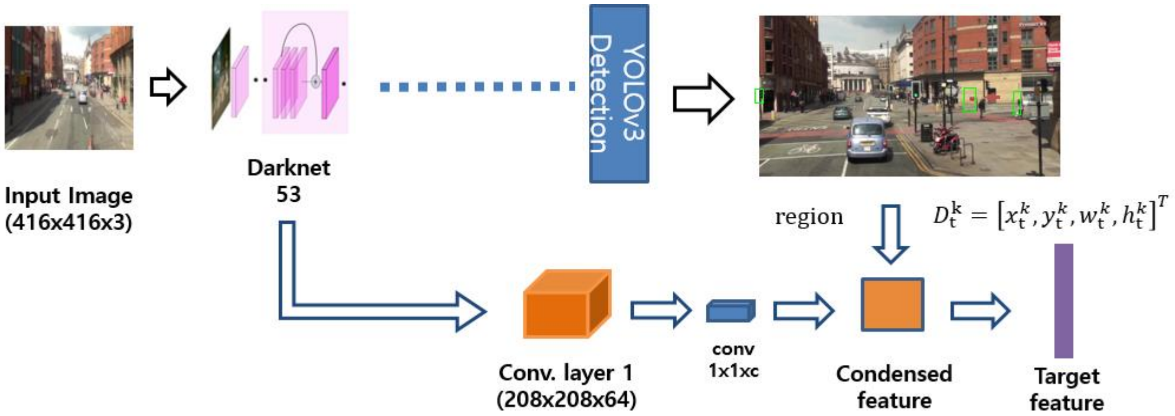


Figure 1. Overall procedure of feature extraction for MPT tracking.

accurate tracking. In this paper, boosted random forest (BRF) [11] is designed as Siamese model structure for data association that is most important part in MPT.

Method

Pedestrian Detection

In real-time application in AV, an online method based on probabilistic inference is applied within the tracking-by-detection framework, where data association between detections and trackers is conducted in every frame. In this study, YOLOv3 (You Only Look Once) [5], a real-time object detection system, is employed as a pedestrian detector. YOLOv3 uses a single neural network to predict the bounding boxes and class probabilities directly from full images in a single evaluation. Because the entire detection pipeline is a single network, the detection performance end-to-end can be quickly and directly optimized.

Appearance Feature Extraction

In vision-based MPT systems, the most important factor in connecting multiple moving pedestrians and trackers is the spatial-temporal feature. However, the temporal feature may degrade performance in moving cameras, so we use only the spatial appearance feature to match pedestrians and trackers. To do this, we extract the condensed feature of the pedestrian by applying the kernel of 1x1xC (channel) to the output feature maps of the second layer of Darknet53, the backbone network of YOLOv3, as shown in Figure 1. Using the proposed feature extraction method, we can effectively reduce feature dimensions and preserve the local information to extract the shape information of the detected target.

Training Siamese Random Forest

For training the proposed SRF tracker, we use output of YOLOv3 feature extractor that pre-trained on ImageNet as the initial parameters of the system. Each feature is handled as following, Eq. (1) makes it possible to transform extracted features into another form of features that having magnitudes of differences on the appearance between the target and detected one. Furthermore, it is possible to infuse flexibility into the model because transformed features represent the objectiveness whether a bounding box contains an object or not.

$$\mathbf{A}\{(a_j)\} = \left\{ abs\left(f^d(x_i) - f^t(x_i)\right) \right\}, \quad i = 1,2,\ldots M \qquad (1)$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ is vector of M (208×208, 43264) dimension feature extracted from fully connected layer of YOLOv3, and f^d, f^t each denotes feature from current detection and tracker. The training data labeled with a scalar as the same object (= 1) or not (= 0). With results of Eq. (1), transformed features, the RF is trained to build an optimal ensemble of decision trees.

The tracker model fills the important role of estimating the most likely location among the previous tracklet to handle appearance changes and the drifting problem. In this study, we applied SRF which has been applied in many computer vision studies, such as object detection, similarity recognition and object tracking [6][7].

In a real driving situation, we have to link detection responses to pedestrians' trajectories on real-time because this is very closely related to pedestrian safety. Therefore, we apply the simple Hungarian algorithm and using a short-term tracklet for hierarchical data association. When a new tracker location is detected, we associate trackers with different candidates using different factors. First, we apply searching window for each of target candidates with a threshold th_w {x,y,w,h} which determined by target's location and size. Then we evaluate the matching score for each tracker-detection pair (t, d) in the search window, combine the observation similarity $Similarity(t|d)$ using SRF, location distance $Dist(t, d)$, and overlap ratio $OR(t, d)$ between tracker(*t*) and detection(*d*). The data association problem is defined by a linear program with the objective function

$$S^* = \arg\max_{d} \left\{ \alpha \cdot Similarity(RF(t,d)) + \beta \cdot {}^{1}/_{Dist(t,d)} + \gamma \cdot OR(t,d) \right\} \qquad (2)$$

where $\alpha, \beta,$ and γ are hyper-parameters that can be adjusted according to the applications. Finally, according to the matching results, the tracker's state is updated by combining the states of the current tracker and the detection information. If the target tracker is overlapped by another object or pedestrian, we can infer whether the target is hidden or not by comparing their bottom location and size of bounding boxes. The update state of the tracker is defined as:

$$\text{state} = \begin{cases} update, & if\ OR(t,d) < \tau_s\ and\ target\ on\ front. \\ keep, & otherwise. \end{cases}$$

In case of the matched target without occlusion or frontal partial occlusion, tracker's state is updated with new information by combining the states of the current tracker and the matched detection information. otherwise, when the target gets full or rear partial occlusion or lost, we keep the scale of the bounding box and tracker information at the last frame and raise its searching window size $th_w\{cx_{t-1}, cy_{t-1}, width, height\}$ for few frame so that the missing pedestrian can be found again. However, if the tracker in this condition does not matched for a given period of time, it is extinguished.

In contrast, if the detections that are not associated with any trackers in the current frame, a new potential tracker should be initialized. This potential tracker is then assigned as a regular tracker if the matching occurs during D (5) times. If the matching count of the potential tracker is less than D times, the potential tracker is assigned a false detection and is eliminated.

Experimental Results

Datasets and evaluation

To experiments the multi-pedestrian tracking performance of the proposed method, we employed the MOT16 Challenge datasets, a common multi-object tracking evaluation reference. It consists of 14 challenging video sequences, seven for each training and testing, comprising with moving and static cameras, crowded scenes, variety viewpoints. The performance test is performed using four video sequences having similar environment with AV.

We use multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP) to evaluate our performance, MOTA accounts for all object configuration errors made by tracker, false positives, misses, mismatches, over all frames. While MOTP is the total error in the estimated position for matched object-tracker pairs over all frames.

All experiments were conducted using a 3.0GHz 8Core CPU and RTX2080ti. Experiments were conducted on three MPT methods, such as Siamese CNN [9], Quadruplet based MOT systems [4] and Dual Matching attention method [10].

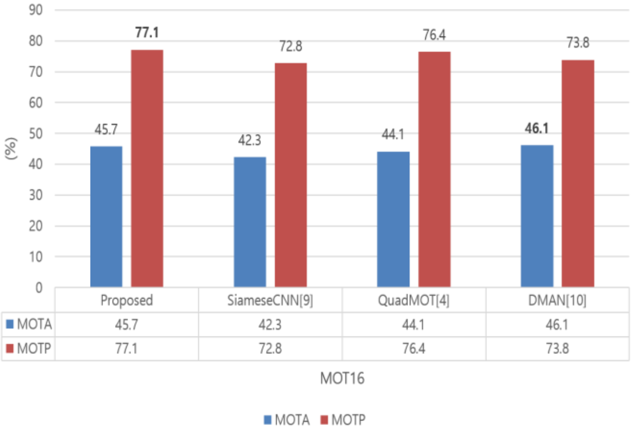


Figure 2. Performance evaluation between four comparable methods in terms of MOTA and MOTP

As shown in the Figure 2, the proposed method shows similar tracking performance with other comparison methods in terms of MOTA, but MOTP shows higher performance from small 0.7% to large 4.3% than other three methods. This means that the proposed method detects the position of the pedestrian more precisely while tracking pedestrian similarly to the comparison method.

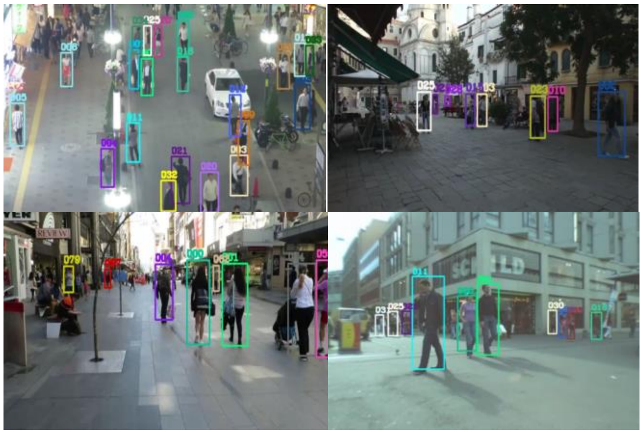


Figure 3. Performance evaluation between four comparable methods in terms of MOTA and MOTP

Figure 3 shows the pedestrian tracking results in various locations with movement camera. From the results, the proposed method shows a relatively accurate tracking result despite camera shake or occlusion. However, if there are still mismatch or tracker ID switching when two pedestrians have a long-term occlusion or similar size and appearance. In terms of speed, the proposed method shows faster processing speed of around 2 or 5 frames per second than the methods using two or more fully connected networks.

Conclusion

In this paper, we proposed a new online MPT algorithm based on SRF tracker model using output features of a YOLOv3. Alternative to the well-known Siamese deep neural networks, the proposed method did not need to run two same Siamese networks at the same time, and similarity distance was measured by BRF. So it could improve the tracking performance.

By using BRF, tracking required much less hyper-parameters then DNN and model complexity could be automatically determined in a data-dependent way. The proposed method had much higher uniform performance in terms of MOTA and MOPT.

In future work, we plan to solve miss tracking caused by the false pedestrian detections, and design light version of BRF for implementing in limited H/W resource. In addition, we need to do more experiments with other studies and various dataset captured in moving vehicles.

Acknowledgement

This work is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1I1A3A01042506) and partially supported by Future Vehicle Leading Technology Development Project funded by Daegu City (DG-2017-01).

References

[1] J. Y. Kwak, B. C. Ko, and J. Y. Nam, "Pedestrian tracking using online boosted Random Ferns learning in far infrared imagery for safe driving at night," IEEE Transactions on Intelligent Transportation System, vol. 18, issue 1, pp. 69-81, Jan., 2017

[2] H. Li, Y. Li, and F. Porikli, "Deep track; learning discriminative feature representations by convolutional neural networks for visual tracking," BMVC, 2014.

[3] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, "Learning by tracking: Siamese CNN for robust target association," CVPR 2017.

[4] J. Son, M. Baek, M. Cho, and B. Han, "Multi-object tracking with quadruplet convolutional neural networks," CVPR 2017.

[5] Joseph Redmon, Ali Farhadi, "YOLOv3: An Incremental Improvement" arXiv:1804.02767, 2018

[6] Lev V.Utkin and Mikhail A.Ryabinin, "A Siamese Deep Forest", arXiv:1704.08715, 2017

[7] L. Bertinetto, J. Valmadre, Joao F.Jenriques , "Fully-Convolutional Siamese Networks for Object Tracking" Knowledge-Based Systems, Vol. 139 Issue C, pp. 13-22, Jan. 2018

[8] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-obj tracking," arXiv:1603.00831 2016.

[9] B. Wang, L. Wang, B. Shuai, Z. Zuo, T. Liu, K. Luk Chan, G. Wang, "Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association", IEEE, 2016

[10] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, M. Hsuan Yang, "Online Multi-Object Tracking with Dual Matching Attention Networks", ECCV, 2018

[11] Y. Mishina, R. Murata, Y. Yamauchi, "Boosted Random Forest". In Proceedings of the International Conference on Computer Vision Theory and Applications (ICCVTA), Lisbon, Portugal, 5–8 January 2014; pp. 594–598.

[12] C. Beyan, A. Temizel, "Adaptive mean-shift for automated multi object tracking", Jour. IET Computer Vision, vol. 6, issue. 1, 2012

[13] X. Li, K. Wang, W. Wang, Y. Li, "A multiple object tracking method using Kalman filter", IEEE International Conference on Information and Automation , 2010

[14] M. Jaward, L. Mihaylova, N. Canagarajah, D. Bull, "Multiple object tracking using particle filters", IEEE Aerospace Conference, 2006

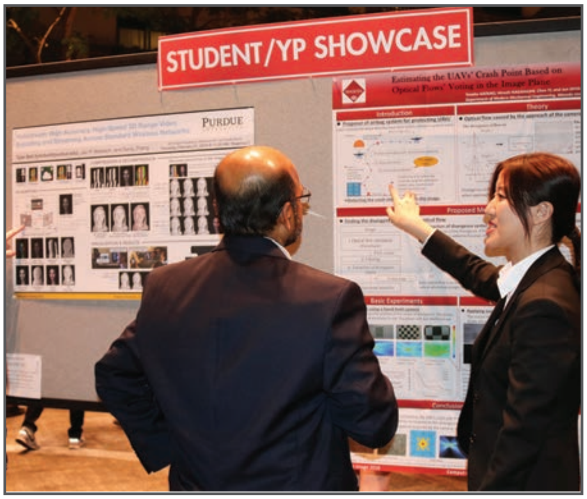
JOIN US AT THE NEXT EI!

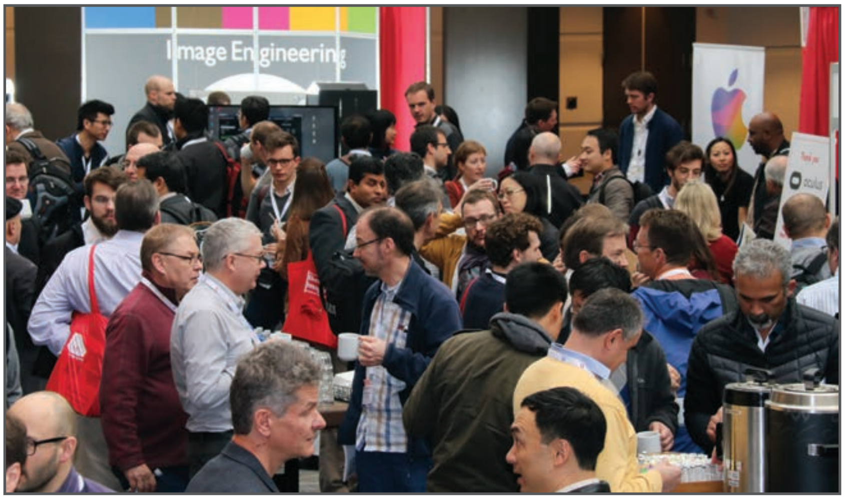
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





AR is a fundamental and transformative interface and platform.
20+ years from my initial working demo to today's AR displays
Optimistic that it won't take another 20 to reach ubiquitous consumer AR!
My first working AR prototype
Today's platforms
Ubiquitous consumer AR

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org