Multi-pedestrian tracking using deep learning technique and tracklet assignment

Mai Thanh Nhat Truong*, Sanghoon Kim*
*Department of Electrical, Electronic, and Control Engineering, Hankyong National University

Abstract

Pedestrian tracking is a particular problem of object tracking, and an important component in various vision-based applications, such as autonomous cars or surveillance systems. After several years of development, pedestrian tracking in videos is still a challenging problem because of various visual properties of objects and surrounding environment. In this research, we propose a tracking-by-detection system for pedestrian tracking, which incorporates Convolutional Neural Network (CNN) and color information. Pedestrians in video frames are localized by a CNN, then detected pedestrians are assigned to their corresponding tracklets based on similarities in color distributions. The experimental results show that our system was able to overcome various difficulties to produce highly accurate tracking results.

1. Introduction

Pedestrian tracking has important roles in many vision-based applications such as traffic monitoring [1], surveillance systems [2], and recently, self-driving car [3]. Generally, the goal of pedestrian tracking algorithms is to locate walking people in video data retrieved from image acquisition devices and to produce a record of the trajectories of the pedestrians, which are called tracklets. Pedestrian tracking, or object tracking in general, can be more difficult because of the variations of object shapes and light conditions, occlusions, sudden change in motions, camera motions, etc. Owing to various difficulties that cannot be solved simultaneously, object tracking methods are usually designed to track objects with specific properties in certain environments [4]. Up to now object tracking have been still a high-complexity and time-consuming task. The complexity is increased when the tracking task is performed in environments with complex surroundings, or the requirement is to track objects with various appearances.

Approaches for object tracking can be categorized into three groups: point tracking, kernel tracking, and silhouette tracking [5]. In point tracking, the target objects are represented by points that contain information of the object properties. Objects are tracked using the relation of the points, and their locations and movements are calculated based on the previous state of these points. However, an external mechanism is required to locate the objects in every frame. Kernel tracking relies on the object shape and appearance, which are called kernels. For instance, the kernel can be a rectangular region or an elliptical shape with an associated histogram. Objects are tracked by calculating the motion of the kernel in consecutive frames. This motion is usually defined as a parametric transformation such as translation, rotation, and affine. In silhouette tracking, the target objects are tracked by their estimated region in each frame. Silhouette tracking approaches use the features extracted from the object region, which are also called models. After having obtained the models, object silhouettes are tracked by using shape matching or contour evolution. Essentially, both these methods can be considered as object segmentation applied in the temporal domain using the object state from the previous frames.

The tracking-by-detection approach is classified as point tracking. In point tracking, the target objects are represented by points that contain information of the object properties. Objects are tracked using the relation of the points, and their locations and movements are calculated based on the previous state of these points. An external mechanism is required to locate the objects in every frame. Several pedestrian trackers using this approach have been proposed recently. Algorithm of Dehghan et al. [6] detects each person using a part-based human detector, then they employ a global data association method based on Generalized Graphs for tracking each individual in the whole video. Lacabex et al. [7] developed a lightweight method for object detection based on background subtraction, window masking and image density projections. The association of detections and trackers is carried out with the Hungarian algorithm on a loss matrix that is constructed according to position, size and appearance. Jiang et al. [8] proposed a system which comprise of a two-stage re-identification algorithm dealing with cases of track drift and re-entry into the field of view individually, in order to match the identities of lost and reappeared targets through a comparison of the affinities between their appearance, size and position, and also to update the status of re-identified targets.

The general disadvantage of tracking-by-detection systems is that, the pedestrian detectors are not reliable because they are constructed by hand-crafted features. The appearance of pedestrians varies significantly when the environment changes. Therefore hand-crafted features are not able to represent the appearance of pedestrians in all possible cases. Moreover, the tracklet association components are also complex, which increases execution time of the tracking system. In this research, we propose a tracking-by-detection system for tracking pedestrian. Our system incorporates two main components: the first component is a pedestrian detector which is a combination of Faster R-CNN and Resnet-101, two CNN-based algorithms. Locations of detected pedestrians from the first component are then fed to the second component, in which spatial overlap properties and color information are used to associate the detected pedestrians to their corresponding tracklets.

2. Proposed method

Our tracking system incorporates two main components: pedestrian detection and tracklet association. The flowchart of our algorithm is illustrated in Figure 1. The pedestrian detector is a combination of Faster R-CNN and Resnet-101, two CNN-based algorithms. This component processes video frames and returns locations of pedestrians in the form of bounding boxes. To associate detected pedestrians to their corresponding tracklets, we use spatial overlap properties and color information to achieve high performance while maintaining low computation cost. The details of each component are presented in the next two subsections.

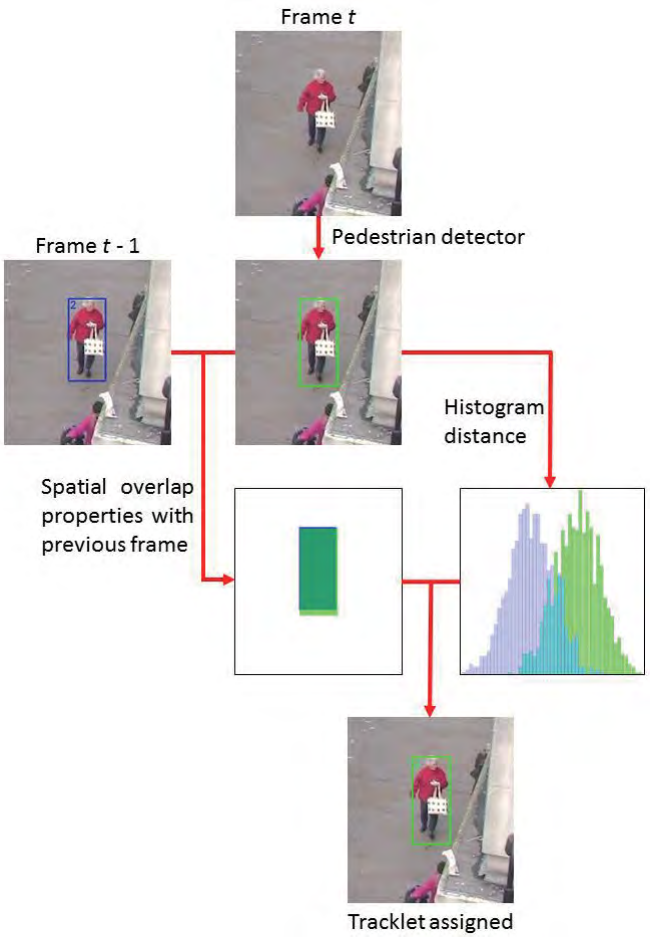


Figure 1. Flowchart of proposed method.

2.1 Pedestrian detection

For pedestrian detection in each video frame, we combine Faster RCNN [9] and Resnet-101 [10]. Resnet-101 is a feature extractor that provides high-level features for Faster R-CNN, an object detector. The combination of Faster R-CNN and Resnet-101 allows the pedestrian detector to produces highly accurate results within reasonable execution time. There have been several CNN-based object detectors and feature extractors, however, several experiments showed that this combination produces the best trade-off between speed and accuracy [11].

Faster R-CNN is a two-stage object detector. In the first stage, images are processed by a feature extractor. In our system the feature extractor is Resnet-101. Features at some intermediate level are used to predict region proposals. In the next stage, these region proposals are used to crop features from the same intermediate feature map which are then fed to the remainder of the feature extractor in order to predict a class and class-specific box refinement for each proposal.

2.2 Tracklet association

After acquiring the detection result from the first component, the second component assigns each detected pedestrian to its corresponding tracklet. Assumes that the video recording device had worked without any interruption and there was no missing frame, we use spatial overlap properties [12] to determine the identity of given pedestrian. The location of the bounding box of a pedestrian is compared to the locations of the other bounding boxes in the previous frame. Because there is no missing frame and the walking speed is slow, the bounding box of a pedestrian in the current video frame always overlap the bounding box of said pedestrian in the previous frame (Figure 2). Hence, the given pedestrian will be assigned to the tracklet that overlapped its bounding box in the current frame. This simple approach allows the tracking system to run in real-time.

If there are more than one bounding box that overlap the region of the given pedestrian, i.e. there were collisions between pedestrians or several pedestrians walking together, we use color distributions to find the correct tracklet. For each bounding box, we calculate 3D histograms in RGB space using 8 × 8 × 8 bins. Then we calculate the statistical distance between two bounding boxes using the Hellinger distance. The pedestrian will be assigned to the tracklet that has lowest distance.

(a) (b) (c)

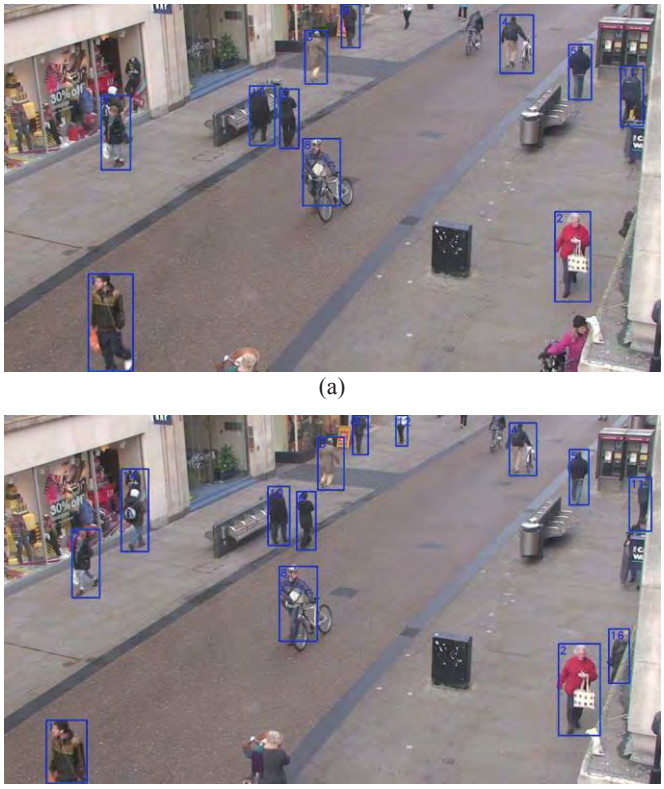
Figure 2. Bounding boxes of a pedestrian in two consecutive frames overlap each other. (a) Bounding box in frame *t* - 1 (b) Bounding box in frame *t* (c) Overlapped region

3. Experimental results

In this section, we present the performance of the proposed pedestrian tracking system. The program was implemented in Python running within Linux operating system. The implementation was deployed in a desktop computer which was equipped with an Intel Core i7-8700K, 32 gigabytes of memory and a Titan XP GPU. OpenCV library was used for processing the video frames and Tensorflow library was used for the implementation of CNN-based pedestrian detector. The video used for testing are acquired from [13]. The results are shown in the following figures. Fig. 3a showed the first frame of the video. In this frame the person that assigned number 7 (top left corner of the video frame) collided with another people. A few seconds later (Fig. 3b), the person number 7 had been still correctly tracked thanks to color information, and the other person that collided was assigned a new number. Other people have been also correctly tracked.

4. Conclusions

Pedestrian tracking is an important component in various vision-based applications, such as autonomous cars or surveillance systems. In this research we proposed a tracking-by-detection system for tracking pedestrian. Our system incorporates two main components: the first component is a pedestrian detector which is a combination of Faster R-CNN and Resnet-101, two CNN-based algorithms. Locations of detected pedestrians from the first component are then fed to the second component, in which spatial overlap properties and color information are used to associate the detected pedestrians to their corresponding tracklets. Resnet-101 is a feature extractor that provides high-level features for Faster R-CNN, an object detector. The combination of Faster R-CNN and Resnet-101 allows the pedestrian detector to produces accurate results in short execution time. Spatial overlap properties and color information yields accurate tracklet assignments while maintaining low computation cost. Experimental results showed the efficiency of the proposed system in pedestrian tracking. For future work, we will improve the performance of our system by improve the accuracy of the object detector, along with proposing a better model for tracklet assignment.



(a)

(b)

Fig. 3. Results from our pedestrian tracking system

5. Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2015R1D1A1A01057518). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] Sutheerakul, C., Kronprasert, N., Kaewmoracharoen, M., Pichayapan, P.: Application of unmanned aerial vehicles to pedestrian traffic monitoring and management for shopping streets. Transp. Res. Procedia 25, 1717-1734 (2017)

[2] Sriram, K.V., Havaldar, R.H.: Human detection and tracking in video surveillance system. In: 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-3. IEEE Press, New York (2016)

[3] Li, F., Zhang, R., You, F.: Fast pedestrian detection and dynamic tracking for intelligent vehicles within V2V cooperative environment. IET Image Process. 11, 833-840 (2017)

[4] Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. 36, 1442-1468 (2014)

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

[6] Dehghan, A., Idrees, H., Zamir, A.R., Shah, M.: Automatic detection and tracking of pedestrians in videos with various crowd densities. In: Weidmann, U., Kirsch, U., Schreckenberg, M. (eds.) Pedestrian and Evacuation Dynamics 2012, pp. 3-19. Springer, Cham (2014)

[7] Lacabex, B., Cuesta-Infante, A., Montemayor, A.S., Pantrigo, J.J.: Lightweight tracking-by-detection system for multiple pedestrian targets. Integr. Comput.-Aided Eng. 23, 299-311 (2016)

[8] Jiang, Y., Shin, H., Ju, J., Ko, H.: Online pedestrian tracking with multi-stage re-identification. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-6. IEEE Press, New York (2017)

[9] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137-1149 (2017)

[10] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. IEEE Press, New York (2016)

[11] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K.: Speed/accuracy trade-offs for modern convolutional object detectors. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3296-3297. IEEE Press, New York (2017)

[12] Bochinski, E., Eiselein, V., Sikora, T.: High-Speed tracking-by-detection without using image information. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-6. IEEE Press, New York (2017)

[13] Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: MOTChallenge 2015: Towards a benchmark for multi-target tracking. arXiv:1504.01942 (2015)