5C-04

A Separated Structure Based Approach for Multiple Object Tracking Robust to Occlusion

Bo CHEN[†1] Toru ABE[†1,†2] Takuo SUGANUMA[†1,†2]
[†1]Graduate School of Information Sciences, Tohoku University
[†2]Cyberscience Center, Tohoku University

1 Introduction

Multiple object tracking (MOT) is a mid-level task in computer vision, it aims to locate multiple objects in the input video, and maintain their identities and trajectories. Existing methods are hard to track objects during occlusion, due to the tracker easily lost occluded object parts. At early stage of research we focus on pedestrian tracking. In this research, we try to solve the occlusion issue with a novel separated structure, which is based on deep learning method. We separate the occlusion scene into foreground and occluded parts, and pursue the former part by an ordinary tracker and the later part by a novel detector, which is trained on occluded images, therefore more robust to occlusion. The validation experiment demonstrated the proposed strategy.

2 Related Work

There are various strategies for handling occlusion issue in MOT [1].

(1) Part-to-whole: Assuming that part of the object is still visible during occlusion, this method divides the object into grids and finds the correspondences between these parts.

(2) Hypothesize-and-test: This method generates occlusion hypotheses based on occludable pair of observations, and tests the hypotheses according to observations at hand.

(3) Buffer-and-recover: This method buffers the state of objects before occlusion, and recovers the states after occlusion based on buffer.

Many works based on above strategies have been proposed. Recently, the deep learning based methods significantly improved the MOT performance. These methods employed CNN based detectors to explore pedestrians in each frame, and then find correspondences between them. They handled the occlusion issue by "Part-to-whole" strategy, but didn't work very well.

3 Proposal

We propose a novel separated structure based approach to solve the occlusion issue.

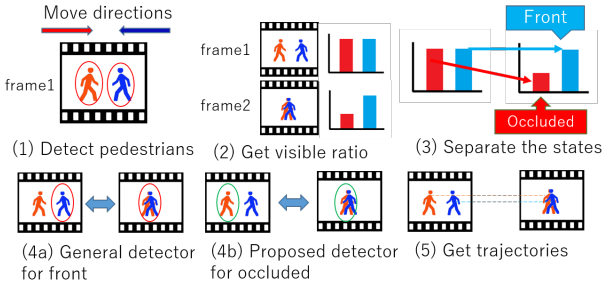

Fig. 1 The flowchart of proposed method.

The assumption of the scenario is: At the beginning there are only complete pedestrians in the frame, they move to each other and the behind one is occluded, then they keep moving and exchange their positions. Figure 1 illustrates the procedure of the proposed method.

(1) The complete pedestrians are extracted by existing detector in each frame. The detector indicates which pixel is pedestrian region.

(2) The visible ratio $VR_i(x)$ of pedestrian x is calculated as:

$$VR_i(x) = \frac{S_i(x)}{S_{max}(x)} \tag{1}$$

where $S_i(x)$ is visible area (pixels) of pedestrian x in frame i, and $S_{max}(x)$ is the max area of the pedestrian x in frame 1, where the pedestrians are not occluded. At the start stage of occlusion, the existing detector still could find slightly occluded pedestrian.

(3) The occlusion scene is separated into foreground and occluded parts by experimentally determined threshold θ for variation of $VR_i(x)$:

$$\begin{array}{l}\text{if } |VR_i(x) - VR_1(x)| < \theta \\ \quad \text{then} \quad \text{foreground} \\ \quad \text{else} \quad \text{occluded}\end{array} \tag{2}$$

In occluded scene, the front pedestrian is "foreground", the back pedestrian is "occluded part".

(4a) The "foreground" is pursued by existing detectors, owing to the sufficient ability for completed objects.

(4b) The novel occlusion-robust detector is applied to "occluded part" when existing tracker failed to pursue.

(5) Finally, the correspondence of both parts is calculated to acquire the trajectories.

The main contribution of this research is the attempt of utilizing occluded images for training of CNN based detector in step (4b). In the existing image data sets used for CNN training, such as ImageNet, the target objects are mostly completed ones, there are rarely occluded objects, which leads to the sensitivity of occlusion for detectors. This fact due to the laborious annotation of images, and enormous cost of manual work. In this research, we try to automatically synthesize the pedestrians to overcome the lack of data.

4 Experiment

The practical detector is trained on large scale data sets, it consists of two parts: (1) Provide position of pedestrian and (2) Classify occlusion state (complete/occluded) of them. Due to lack of data, it's hard to evaluate our idea by detector, we evaluated the idea by a simple occlusion state classifier on small-scale data set [2] .

4.1 Synthesis of occluded images

The data set contains three parts: Original pedestrian image, label of the pedestrian positions, and masks showing the pedestrian region. Figure 2 shows example of data set.

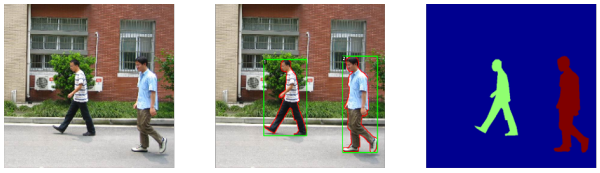

Fig. 2 Example of data set.

We extracted the individual pedestrians, and moved two pedestrians to each other on horizontal directions to synthesize occlusion scenes. We simplified the situation because the actual occlusion scene may contain complex movements and various amount of pedestrians. Figure 3 shows the synthesis process of occluded images.

4.2 Experiment setting

We evaluated the proposed strategy on synthesized data set. We tuned a ResNet [3] based classifier and trained it on two data sets: (1) Original complete images (156 images), and (2) 4104 synthesized images which generated from original images, 3604 for training, and 500 for test.

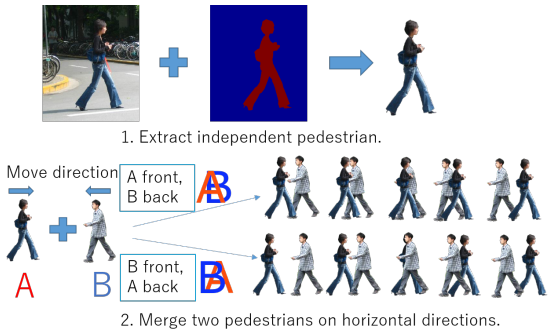

Fig. 3 Synthesis of occluded images.

4.3 Result and Analysis

We compared the two classifiers on the 500 occluded synthesized images, The result is shown in Table 1.

Table. 1 Experiment result

Classifier	Accuracy
(1) trained with 156 comp. images	0.957
(2) trained with 3604 syn. images	0.981

Here the accuracy is how much occluded images were correctly classified. The result demonstrated that the classifier will improve the accuracy for occluded scene by using more occluded images for training.

5 Conclusion

In this research, we proposed a novel separated structure for MOT, which divides the occlusion into foreground and occluded part, and pursues the later part by a new detector which is trained on occluded images. We verified our idea on small scale data set, the experiment result demonstrated it.

References

[1] Luo, W. et al.: Multiple object tracking: A literature review, *CoRR*, Vol. abs/1409.7618 (2014).

[2] Wang, L. et al.: Penn-Fudan Database for Pedestrian Detection and Segmentation, `https://www.cis.upenn.edu/~jshi/ped_html/`. (accessed 2018-12-18).

[3] He, K. et al.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016).