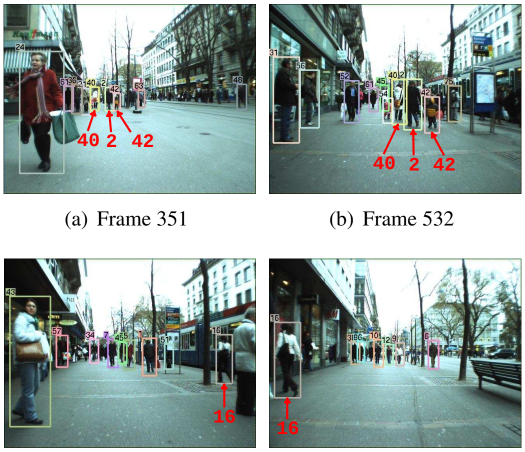
Tracklet Association in Detect-then-track Paradigm for Long-term Multi-Person Tracking
(Extended Abstract)

Bing Wang, Gang Wang, Kap Luk Chan, Li Wang
School of Electrical and Electronic Engineering, Nanyang Technological University
50 Nanyang Avenue, 639798, Singapore
{wang0775,wanggang,eklchan,wa0002li}@ntu.edu.sg

1. Introduction

Recently, significant progress has been reported in human detection and tracking, based on the popular tracking paradigm: detect-then-track. A typical way to doing this is to track multiple targets frame by frame, which often encounters irrecoverable errors if a target is undetected in one or more successive frames or if two detections are erroneously linked. To overcome this weakness, the global trajectory optimization methods over batches of frames have been proposed in recent years. These methods are often based on graphical network optimization in which the the nodes are represented by detection responses. Such methods often fail to handle the problems of long-term tracking in crowded scene well. To alleviate this, some researchers try to use the track fragments (tracklets) as graph nodes. However, due to the long duration gaps between tracklets and less effective appearance-based models, many existing methods are not capable of handling long-term occlusions and interactions between targets.

In [4], we present a novel introduction of online target specific metric learning in track fragment (tracklet) association by network flow optimization for long-term multi-person tracking. Instead of detection responses, tracklets are used as the nodes in the network graph, with edges defined by a cost computed from a novel tracklet affinity scores. In the proposed framework, we learn target-specific metrics so that target-specific properties can be efficiently explored for more discriminative models. Our learning is online throughout and our target-specific metrics are adaptive to local segments. The proposed framework allows longer gaps between tracklets to be linked, which make it more capable of handling long-term occlusions and interactions between targets.



(a) Frame 351 (b) Frame 532

(c) Frame 673 (d) Frame 877

Figure 1. Frames from ETH dataset with target identities labeled by the method of [4]. The ID labels 2, 40, 42 in the top row and ID label 16 in the bottom row remain unchanged after many occlusions and interactions over more than 180 frame intervals.

2. Tracklets Association Framework

The framework proposed in [4] is shown in Figure 2. Given a video input, we first detect pedestrians in each frame by an existing detector. The initial tracklets are generated based on motion trajectory using successive shortest path algorithm [2]. Nevertheless, the initial tracklets may be unreliable because the detection responses in one tracklet may come from more than one person. Hence, we use the online learned target-specific metrics to refine these initial tracklets for reliable tracklets. The cost-flow network formulation is based on the reliable tracklets and network flow optimization yields the long-term trajectories of multiple persons. Estimating the transition costs is the key factor in the min-cost network flow optimization. We propose to learn target-specific segment-wise appearance-based model online for estimating the transition costs.

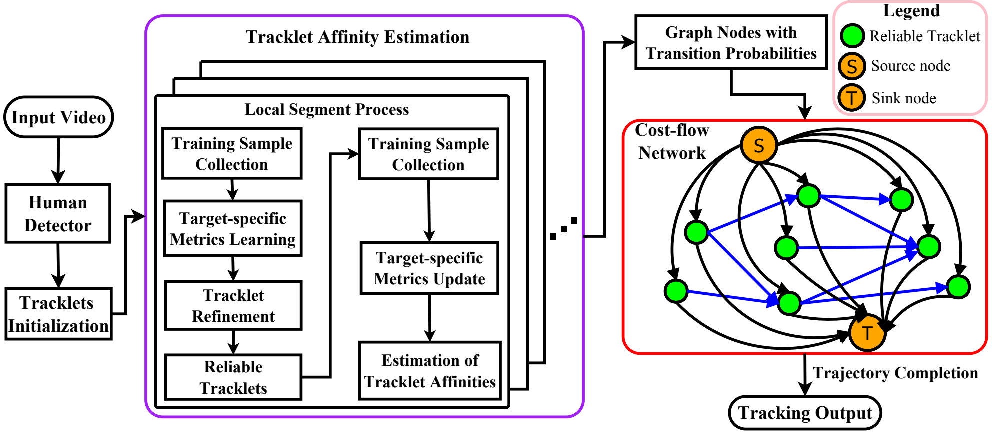
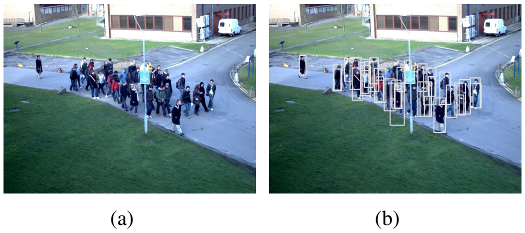

Input Video
Human Detector
Tracklets Initialization
Tracklet Affinity Estimation
Local Segment Process
Training Sample Collection
Target-specific Metrics Learning
Tracklet Refinement
Reliable Tracklets
Training Sample Collection
Target-specific Metrics Update
Estimation of Tracklet Affinities
Graph Nodes with Transition Probabilities
Legend
Reliable Tracklet
S Source node
T Sink node
Cost-flow Network
S
T
Trajectory Completion
Tracking Output

Figure 2. The proposed framework. In the cost-flow network, each node denotes a reliable tracklet; The flow costs of edges are defined by negative log of the affinity scores, which are obtained through a two-step target-specific metric learning and metric refinement processes on segments of short-time sequences known as local segments.

3. Pedestrian Detection with Occlusion Handling

Pedestrian detection is a crucial first step in the detect-then-track approach. Human detection methods based on HOG and DPM have been commonly used. We used the latter in the current work because the mixture of deformable part models can handle partial occlusion better. Since then, we have further improved the detection in the part-based detector for crowded scenes as shown in Figure 3. We propose an effective detection framework to handle more severe occlusions in highly crowded scenes, we reformulate the score computation of body parts in the generic deformable part-based models [1] and utilize the online learned dictionary to refine the detection responses.

In our previous work [5], we present a pedestrian detection method by using the depth data obtained from 3D imaging methods (see Figure 4). This helps when the ambiguity of detection from 2D images cannot be resolved.


(a)

(b)

Figure 3. An example of the detection results from [3].

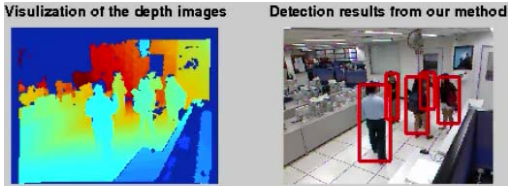

Visulization of the depth images
Detection results from our method

Figure 4. An example of the detection results from [5].

4. Conclusions

In this article, we report our recent works for long-term multi-person tracking. As we can see in Figure 1, our proposed framework is capable of handling long-term multi-person tracking problems. Improving human detection in crowded scene will help the tracking. We further improved the performance of part-based detector in [3]. We also explored the use of depth sensing in [5] to resolve ambiguity in detection from 2D images.

References

[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010. 2

[2] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In CVPR, 2011. 1

[3] B. Wang, K. L. Chan, G. Wang, and H. J. Zhang. Pedestrian detection in highly crowded scenes using online dictionary learning for occlusion handling. submitted to ICIP, 2014 (under review). 2

[4] B. Wang, G. Wang, K. L. Chan, and L. Wang. Tracklet association with online target-specific metric learning. In CVPR, 2014. 1

[5] L. Wang, K. L. Chan, and G. Wang. Human detection with occlusion handling by over-segmentation and clustering on foreground regions. In ACCV Workshops, 2012. 2