Neurocomputing 139 (2014) 220–232

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Single/cross-camera multiple-person tracking by graph matching

Weizhi Nie [a], Anan Liu [a,*], Yuting Su [a], Huanbo Luan [c], Zhaoxuan Yang [a], Liujuan Cao [b], Rongrong Ji [b]

[a] School of Electronics Information Engineering, Tianjin University, 300072, China
[b] School of Information Science and Engineering, Xiamen University, 361005, China
[c] School of Computing, National University of Singapore, Singapore

ARTICLE INFO

*Article history:*
Received 22 August 2013
Received in revised form
10 December 2013
Accepted 5 February 2014
Communicated by Ran He
Available online 13 April 2014

*Keywords:*
Graph matching
Affinity constraint
Part-based model
Object tracking
Cross camera

ABSTRACT

Single and cross-camera multiple person tracking in unconstrained condition is an extremely challenging task in computer vision. Facing the main difficulties caused by the existence of occlusion in single-camera scenario and the occurrence of transition in cross-camera scenario, we propose a unified framework formulated in graph matching with affinity constraints for both single and cross-camera tracking tasks. To our knowledge, our work is the first to unify two kinds of tracking problems with the same framework by graph matching. The proposed method consists of two steps, tracklet generation and tracklet association. First, we implement the modified part-based human detector and the Tracking-Modeling-Detection (TMD) method for tracklet generation. Then we propose to associate tracklets by graph matching which is mathematically formulated into the Rayleigh Quotients Maximization. The comparison experiments show that the proposed method can produce the competing results with the state-of-the-art methods.



## 1. Introduction

Multiple-person tracking in unconstrained condition is a popular research area in computer vision due to its wide applications in intelligent video surveillance, human computer interaction, and so on. There are a few factors making this problem extremely challenging. For example, partial occlusion in single camera scenario might cause the pseudo-disappearance of the tracking target and view transition in cross-camera scenario would miss the correspondence between multiple persons if no prior knowledge is provided. Over the past two decades, a large number of trackings [1–4] have been proposed to handle certain aspects of these challenges. These methods can be roughly classified into two kinds: (1) Short-term tracking: the representative approaches, like Meanshift, Kalman and Particle Filter [5–7], are widely used to perform frame-to-frame tracking. This kind of trackers often fails in some special scenarios because there is no decision-making mechanism to decide whether the targets disappear or are occluded. (2) Long-term tracking: this kind of methods firstly implements a detector to localize the object in each frame. Then it utilizes the tracker to associate the detection results. Since this method can directly address the post-failure behavior by recovering the object which disappears by occlusion, it can be utilized for long-term tracking. One representative approach is tracking-by-detection. Shu et al. [8] trained a decision model with human body part information for occlusion handling and implemented a first-order Markov for trajectory generation. Similarly, Andriluka et al. [9] focused on developing the robust object detector with discriminative visual features and took advantages of both detector and tracker for data association.

With the increasing interests from both academy and industry, cross-camera multiple-person tracking has become a hotspot [10,11]. Typically, cross-camera tracking consists of two steps, object tracking in single camera and data association between multiple cameras. Consequently, besides the difficulty in single-camera tracking, the main challenge of cross-camera tracking is to associate the trajectories of human movement in different views. Li et al. [12] combined the detector and the MoSIFT-based tracker to handle the multiple-camera person tracking in a nursing home based on the prior knowledge of spatial relationship and the heuristic data association methods for different kinds of transitions among multiple cameras. Based on Li's method [12], Zhong et al. [13] brought in temporal information to associate individual human trajectories from different views to improve the accuracy of cross-camera tracking. The crucial problem caused by cross-camera tracking lies in the drastically increasing data. The high computation complexity might be solved by the promising GPU or many-core computing [14].

Data association methods play an important role in handling occlusion in single-camera scenario and transition in cross-camera

* Corresponding author.
*E-mail address:* anan0422@gmail.com (A. Liu).

http://dx.doi.org/10.1016/j.neucom.2014.02.040
0925-2312/

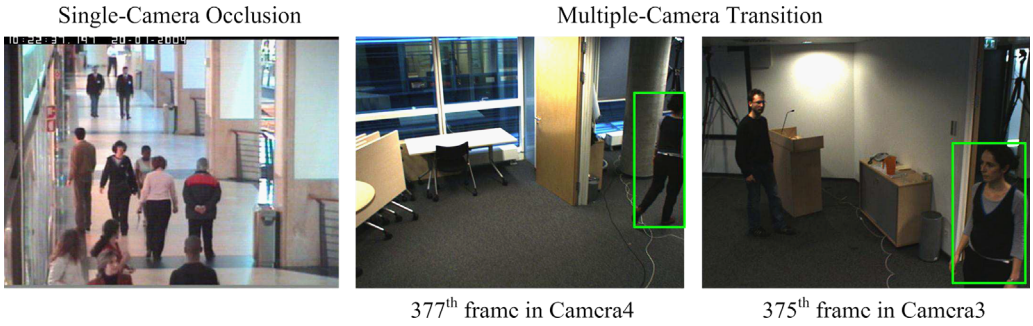


**Fig. 1.** Samples of difficulties in multiple-person tracking. Left: occlusion sample in single-camera scenario from Caviar dataset; right: transition sample in green box in cross-camera scenario from ICPR 2012 Contest dataset. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

scenario (Fig. 1) for long-term trajectory generation [15–17]. In order to deal with these problems, we propose a unified framework for both single and cross-camera multiple-person tracking. The proposed method consists of two steps, tracklet generation and tracklet association. For tracklet generation, we implement a modified part-based human detector to localize the persons for initialization and then utilize the Tacking-Modeling-Detection (TMD) method to track individuals. For tracklet association, each tracklet can be considered as a node in the XY-T space and the tracklet nodes, before and after occlusion in single-camera scenario or transition in cross-camera scenario, compose two graphs respectively. Consequently, we unify two seemingly different kinds of tracklet association into the same framework of graph matching based on both visual similarity and spatiotemporal context. The main contributions lie in three aspects:

- We propose a unified framework for both single and cross-camera multiple-person tracking. To our knowledge, our work is the first to unify two kinds of tracking problems with the same framework.
- The tracklet association problem is converted into graph matching with affinity constraints and is further mathematically formulated into the Rayleigh Quotients Maximization.
- We leverage both person-wise and part-wise attributes for similarity measurement between tracklets to overcome the uncertainty and noise for feature representation and matching induced by complex background and dynamic appearance.

The comparison experiments show that the proposed method can produce the competing results with the state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 will introduce the state-of-the-art methods on multiple-person tracking which are closely related to this paper. Section 3 will illustrate the system overview. The tracklet association via graph matching will be detailed in Section 4. The experimental method and results will be respectively presented in Sections 5 and 6. At last, we conclude this paper.

## 2. Related work

The recent work on multiple-person tracking usually takes advantages of both the detector and the tracker for long-term tracking [18,19]. These methods mainly contain three key techniques, object detection, short-term tracking for tracklet generation, and tracklet association for long-term trajectory generation.

In the stage of object detection, many methods have been developed to get accurate detection results. First, background substraction is usually implemented to filter out the regions which have low motion intensity and consequently cannot contain any human [20–22]. Then different detectors can be implemented on the motorial foreground regions for human localization. Dalal and Triggs [23] proposed the Histogram of Oriented Gradient-based SVM classifier for human detection. Although it generally works well in human detection with discriminative shape feature, it usually fails when occlusion happens. Shu et al. [8] learned a part-based person-specific SVM classifier, which captures articulations of human body, to deal with partial occlusions. To improve detection results, Yang and Nevatia [24] proposed to train a classifier with appearance characteristics to predict the potential positions of the person and integrate both detection and prediction results for refinement.

The goal of short-term tracking is to generate reliable tracklet. The traditional tracking methods, including Meanshift [5], Kalman [6] and Particle Filter [7], have been used for this step. However, these methods often fail in the case of the occurrence of occlusion because the occluded object cannot be easily recovered whenever it is missing. Therefore, lots of methods focus on designing the decision-making mechanism to decide whether the target has disappeared and how to recover it [2,25,26]. This kind of method usually requires a large amount of training sets to improve the performance of model learning [27,28]. However, all these methods strictly separate the training and testing phases which means that the appearance variations which are not covered by the training set will never be fed into the model and therefore the performance of the model cannot be boosted adaptively. To handle this problem, Kalal et al. [29] proposed a Tracking-Modeling-Detection algorithm in an adaptive manner for long-term tracking. Inspired by online learning, this method can learn a classifier for each target and then apply this classifier for object detection. The learnt detector can re-initialize the tracker whenever the target gets lost.

The goal of data association is to link the tracklets to form complete trajectories for individual object. Traditional multiple object tracking methods, such as multi-hypothesis tracking [30] and joint probabilistic data association filters [31], jointly consider the data association from sensor measurements to multiple overlapping tracks. However, they can model only a few state transitions due to the exponentially increasing complexity and have difficulty in utilizing physical exclusion constraints between objects for modeling. For the solution of objective function, Jiang et al. [32] employed the integer linear programming to associate the fixed number of tracklets. To relax the prior of fixed number of tracklets, Berclaz et al. [33] introduced the virtual source and sinked locations to initiate and terminate trajectories. This method can only achieve satisfying performance in constrained condition when no special cases, like person entrance, exit and occlusion, occur. The advanced techniques are required for tracklet association in an unconstrained environment to generate robust and 'lifetime' trajectories.