Multi-Person Tracking-by-Detection based on Calibrated Multi-Camera Systems

Xiaoyan Jiang, Erik Rodner, and Joachim Denzler

Computer Vision Group Jena
Friedrich Schiller University of Jena
{xiaoyan.jiang,erik.rodner,joachim.denzler}@uni-jena.de
http://www.inf-cv.uni-jena.de

Abstract. In this paper, we present an approach for tackling the problem of automatically detecting and tracking a varying number of people in complex scenes. We follow a robust and fast framework to handle unreliable detections from each camera by extensively making use of multi-camera systems to handle occlusions and ambiguities. Instead of using the detections of each frame directly for tracking, we associate and combine the detections to form so called tracklets. From the triangulation relationship between two views, the 3D trajectory is estimated and back-projected to provide valuable cues for particle filter tracking. Most importantly, a novel motion model considering different velocity cues is proposed for particle filter tracking. Experiments are done on the challenging dataset PETS'09 to show the benefits of our approach and the integrated multi-camera extensions.

1 Introduction

Multi-object tracking is important for various applications in computer vision, such as visual surveillance, traffic control, sports analysis, or activity recognition. Since cameras are getting cheaper and tracking definitely benefits from organized multiple cameras with different views, it is sensible to track multiple objects based on calibrated multi-camera systems. In general, tracking multiple objects in real-time in an accurate way is very challenging due to background clutter, occlusion between objects and background, and the appearance similarity between objects to be tracked. The difficulties additionally arise from the aspect of how to reliably fuse the information from individual cameras and how to perform robust global tracking in an efficient manner.

With the improvement of detection algorithms [1,2] both in accuracy and computational feasibility, tracking-by-detection is one of the most popular concepts for tracking [3,4]. Targets to be tracked can be initialized by continuously applying detectors to single image frames. Typically, the output of a detector is a set of image regions with confidence scores. Incorporating temporal context here is necessary due to the high amount of false positives and missing detections as shown in the left part of Fig. 1. Recent tracking approaches [3,4,5,6] try to associate the detections and track objects from uncalibrated single cameras.

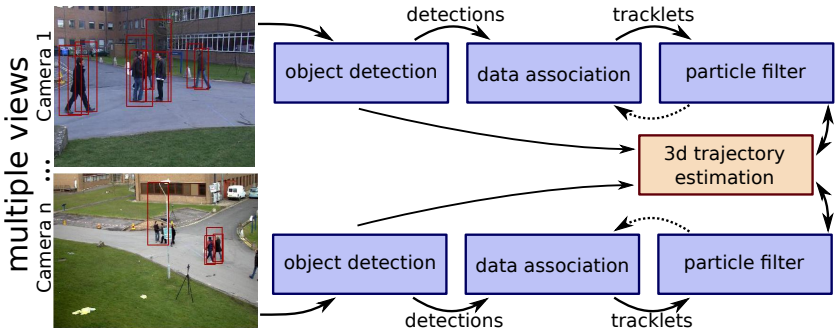

Fig. 1. Overview of our framework. Left images show outputs of a person detector [1] from two views of the PETS'09 database highlighting the necessity of tracking and 3D reasoning.

Additionally, multi-camera systems are also used for efficient tracking. Correspondences of observations of walking humans across multiple cameras can be established by geometric constraints like a planar homography [7]. The work of [8] reconstructs the top-view of the ground plane and map the vertical axes of a person in each view to the top-view to intersect at a single point that is assumed to be the location of the person on the ground. Many other papers also make full use of the ground plane assumption [9,10].

We intend to utilize two or more calibrated cameras and fuse the information in a joint tracking-by-detection framework without using homography restrictions or top-view images. As we will see later, this yields more precise results. In details, detections in single camera images with lower confidence scores which are considered to belong to the same objects are rejected first. Afterwards, detections are associated to form more reliable tracklets. Once tracklets are found, they are used to update the motion model as well as the target model for particle filter tracking. Secondly, global tracking based on estimating the 3D position is realized, where we focus primarily on occlusion reasoning from a geometrical point of view. Thirdly, particle filters are initialized with tracklets and the motion is estimated in subsequent frames.

The paper is structured as follows. The details of our algorithm are presented in Section 2. Experiments on PETS'09 [1] are evaluated qualitatively and compared to other state-of-the-art algorithms in Section 3. Finally, Section 4 concludes the paper with a summary and an outlook.

2 Tracking-by-Detection with Multi-Camera Systems

A block diagram of our tracking approach is shown in Fig. 1. The algorithm mainly consists of three parts: data association, particle filters for tracking, and 3D trajectory estimation. First of all, we use data association techniques to

[1] `http://www.cvg.rdg.ac.uk/PETS2009/`

combine detections after k consecutive frames to form tracklets. Each tracklet is then associated with a particle filter. The key idea is that the particle filter additionally uses the estimated 3D trajectory to fuse the information from multiple cameras. Furthermore, back-projecting this 3D trajectory into every view allows for recovering from missed detections. To eliminate duplicate tracking results, we use several similarity measurements based on appearance features as well as geometry reasoning.

2.1 Data Association

Initializing a particle filter tracker for each detection directly in each frame may lead to unreliable tracking results. Therefore, we use an intuitive matching criteria to find corresponding detections in k subsequent frames. If such a correspondence is found, the detections in all k frames define a tracklet and are taken into account for particle filter tracking. This reduces the number of false positive detections to a large extent. The correspondences are found by greedy association, which showed results comparable to the assignment problem solved by the Hungarian algorithm [3]. We basically follow the work of [3], but modify it in two important aspects: first, we associate the detections in single cameras to get more reliable tracklets. Furthermore, a calibrated multi-camera system is used to estimate the 3D trajectory and use the projections in each camera to provide valuable cues especially in the case of occlusions and ambiguities.

To find the best associated detections in time t in camera i, we consider the Manhattan distance and the overlap ratio between each current detection d with the previous detection from a tracklet T using a gating function:

$$g(d,T) = \begin{cases} 1 & \text{if } o(d,T) \geq \sigma \text{ and } M(d,T) \leq \xi(T) \\ 0 & \text{otherwise} \end{cases} \quad . \tag{1}$$

The value $g(d,T) = 1$ indicates that this detection belongs to the tracklet considered. In equation (1), the overlap ratio $o(d,T)$ between the two regions is defined as follows:

$$o(d,T) = 2 \cdot \frac{d \cap T}{|d| \cup |T|} \quad , \tag{2}$$

where we interpret d and T as sets of image pixels. The parameter σ is set experimentally. The Manhattan distance $M(d,T)$ is thresholded depending on the size of the tracklet:

$$\xi(T) = \alpha \cdot [\text{height}(T) + \text{width}(T)] \quad , \tag{3}$$

where the parameter $\alpha < 1$ is manually chosen.

If a detection passes the gating function, it will be associated to the corresponding tracklet. Furthermore, the input to the thresholding operation is a set of ranked detections of which the detections with higher scores that satisfy the conditions stated above will be selected primarily.

2.2 Data Fusion from Multiple Views

Fusing information from multiple views is done by reconstructing the 3D position of the centroid of the object. With the knowledge of epipolar geometry, it is known that a corresponding pair of points in two cameras is limited to epipolar lines [11]. The correspondence of detections is obtained by the combination of Euclidean distances between their centers to respective epipolar lines and appearance similarity between them. Afterwards, we estimate the 3D position of the object center from two views using triangulation [11].

3D Trajectory After obtaining all possible 3D points from all the cameras, 3D points which are near to each other are considered to belong to the same object, which are merged by simple averaging of their 3D vectors. The 3D trajectory of an object is formed based on associating the 3D points frame by frame.

Occlusion Reasoning The most challenging part of multi-object tracking is how to tackle tracking under occlusion. When there exists inter-object occlusion or the object is occluded by background objects, the target in some views will be partially (or totally) invisible. This might lead to the failure of both detection and tracking. Therefore, many works consider occlusion reasoning to improve the results [3]. The inter-object occlusion reasoning is done by considering the intermediate detection confidence as a part of the observation model of particle filters. In [12], occlusion is taken into account by calculating the visible parts of the object and trajectory estimation is achieved by energy function minimization.

However, we want to take advantage of multiple views and intend to perform inter-object occlusion reasoning from a geometrical point of view. We assume two kinds of occlusion: first situation, if an object was trackable previously and there is another one or more trackable objects nearby, then we regard this object as occluded by another object, no matter how much proportion of the object is invisible; second situation, if one detection in a view has more than one corresponding detections in other views for several frames, then it is supposed that there is occlusion between these detections. After obtaining tracklets and the 3D trajectories of the objects, they are used to update the target model and the motion model of the particle filter, which is explained in the next section.

2.3 Kernel based Particle Filter

In case of clutter environments, the assumption of Gaussian distributed object states does not hold. In contrast, particle filters are able to model flexible and multi-modal distributions, and are due to this reason better suited for tracking than Kalman filters [13]. Particle filters use a set of weighted samples, referred to as particles, to model the posterior distribution [13]:

$$\chi_t = \{\mathbf{x}_t^{[1]}, \mathbf{x}_t^{[2]}, ..., \mathbf{x}_t^{[M]}\} \ , \tag{4}$$

where M is the number of particles which is constant in our case and $\mathbf{x}_t^{[m]}$ (with $1 \leq m \leq M$) is a hypothesis of the 2D state (x, y, hx, hy) (object center and half of the width and height of the object) at time t.

Initialization and Termination Initialization of trackers is done using two different sources: the newly assigned tracklets and the projections of 3D trajectories of the object which are within detections. We also check whether these cues have not been associated with any existing tracker. Initially, particles are distributed uniformly over the initial region with the same weight $1/M$.

The appearance model of a target or a candidate is a kernel-based RGB histogram [14]. The kernel assigns higher weights to the samples close to the target region centroid to reduce the effect of peripheral samples, which might be affected by occlusions from the background. After initialization, particles will be propagated to the new hypothesis states according to the motion model.

Propagation In some papers like [3], people are assumed to walk with constant velocity and the motion model is empirically configured in advance. This assumption is not valid, when people stop walking during a period of time or increase speed. Furthermore, this leads to particles, which may converge to totally wrong positions or scale because of ambiguities in the scene. Therefore, we decide to utilize a robust motion model to guide the particles.

The motion model is composed of three different velocities:

$$\mathbf{v} = \beta \cdot \mathbf{v}_T + \eta \cdot \mathbf{v}_O + \gamma \cdot \mathbf{v}_S \quad \text{with} \quad \beta + \eta + \gamma = 1 \tag{5}$$

where $\mathbf{v}_T$ is the velocity of the associated tracklet at the current time step t, $\mathbf{v}_O$ is the velocity of the tracker at previous time step $t-1$, and $\mathbf{v}_S$ is the velocity of the back projection from the corresponding 3D trajectory at current time step t. In normal situations β, η, and γ are set equally. During occlusion, the detections associated with a tracker in a single view are considered to be unreliable. Thus, β is set to be lower, while γ is given higher weight to incorporate useful information from other cameras. Besides that, $\mathbf{v}$ should not be greater than a maximum velocity. For people who are walking, this maximum velocity could be defined by twice of the normal speed of a person. Otherwise, $\mathbf{v}$ equals to the previous velocity to reduce abrupt movement. One advantage of utilization of this motion model is that the particles can still propagate correctly even during occlusion by fusing useful cues from other cameras.

Observation Given the propagated particles, the weights of individual particles are obtained by the Bhattacharyya coefficient [14] between the candidate and the target. Most importantly, the target model is updated by the associated detection which is totally visible and satisfies object appearance consistency. Appearance consistency is based on the reasonable assumption that the appearance of a person in two consecutive frames does not change significantly.

The final state is estimated by the combination of the mean and the maximum of the modeled posterior distribution:

$$\mathbf{x}_t = \frac{1}{2}\left(\sum_{m=1}^{N} w_t^{[m]} \cdot \mathbf{x}_t^{[m]} + \underset{m=1\ldots N}{\arg\max}\ \mathbf{x}_t^{[m]}\right) \ , \tag{6}$$

where $\mathbf{x}_t$ is the final state at time t and $w_t^{[m]}$ is the corresponding weight of the particle $\mathbf{x}_t^{[m]}$.

3 Experiments

Datasets There are not many publicly suitable datasets for multi-person tracking in multi-camera systems. Since PETS09/S2.L1 contains different types of human movements, different types of occlusion, cameras located in different angles with different illumination and different resolutions, we choose this challenging dataset for testing our algorithm.

In all our experiments, we do not train the detector specifically for this dataset and do not perform background extraction. We use the detector of [1] and the corresponding source code and learned models. Furthermore, no special information of this dataset is used, which allows for applying our algorithm to other scenarios and datasets. The ground truth of the first view was provided by Anton Andriyenko[2].

Experimental Setup We use the CLEAR MOT metrics [15] and the evaluation program of [6] to analyze the tracking performance. We compare our algorithm with the state-of-the-art results on PETS'09 S2.L1 in Table 1, where the performance of other methods are provided in [3]. MOTP considers the precision of estimated positions and MOTA takes misses, false positives, and mismatches into account [15]. The parameter k in this dataset is set to 3 and we use 1000 units in the world coordinate system [6] as a threshold to recognize 3D points belonging to the same object. Furthermore, 8 bins per color channel are used to compute the RGB histograms and σ is set to 0.4. $M = 100$ particles are used in one tracker.

Evaluation Fig. 2 shows six sample images from the dataset. Different colors identify the different objects with corresponding 2D trajectories. From the images, we can see that most of the people can be tracked correctly even during occlusion. There are sometimes double or more assigned trackers to the same object, partially arised from initialization of trackers by back projected points of 3D positions that within detections or from missing detections that may separate long tracklets into several shorter tracklets. The precision of multi-camera system calibration has heavy effect on the final results, since the 3D positions of the objects has large impact especially during occlusion.

Evaluation results are shown in Table 1. We can see that our approach outperforms other methods with respect to the MOTP value. This is mainly due to the utilization of cues from multiple cameras, which is especially useful because of missing detections in the first view. Compared to other state-of-the-art methods, the MOTA value of our approach is lower. This is mainly caused by reassignment of the same objects or the model update of the particle filters when groups of people split and merge again. We plan to overcome these issues by not

[2] `http://www.gris.informatik.tu-darmstadt.de/~aandriye/data.html`

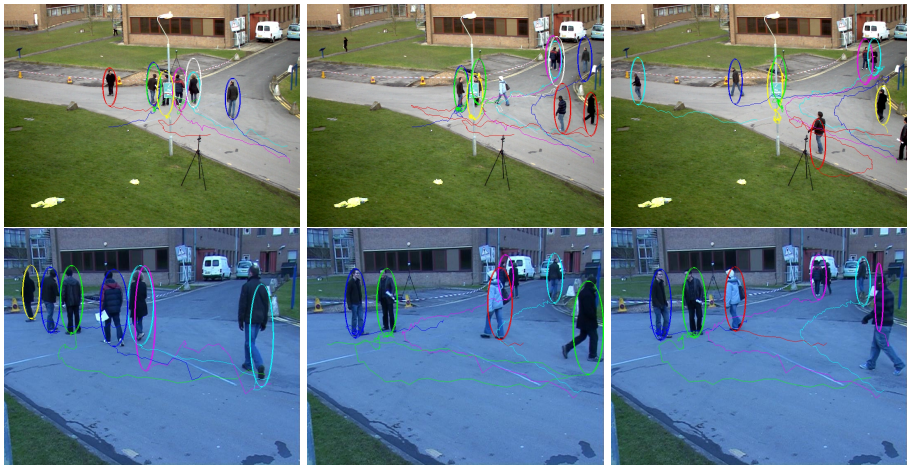

Fig. 2. Tracking results: top and bottom row show results in view 1 and 5, respectively.

Table 1. Results of our approach on PETS'09/S2.L1 compared to state-of-the-art.

Algorithm	MOTP	MOTA
Our approach	**78.8%**	60.8%
Breitenstein et al. [3]	56.3%	79.7%
Yang et al. [16]	53.8%	75.9%
Berclaz et al. [9]	60.0%	66.0%
Andriyenko et al. [6]	76.1%	**81.4%**

only considering the centroid of tracked objects but also the complete 3D shape to allow for a more exact data association.

Runtime Performance The entire system is implemented in C++, except that the detection is the MATLAB source code of [1], without taking advantage of GPU processing. For tracking without considering the time used for detection, the average runtime is 2.6 seconds for each time step processing images from 7 cameras. Time measurements were done on a standard PC with Intel Core i5 2.8GHz processor.

4 Conclusion

In this paper, we proposed a tracking-by-detection framework for multi-camera systems. We showed that estimating the 3D position of the tracked objects can help to solve for ambiguities and provides more robustness to occlusions. Our framework is based on an efficient detection algorithm, several intuitive rejection rules, and the combination of several appearance as well as geometry cues. Specifically speaking, the performance during occlusion is improved by integrat-

ing cues from multiple cameras. This is reflected in the novel motion model incorporated in particle filters, where the importance from 3D trajectory is higher during occlusion.

For future research, we plan to integrate more complex motion models and uncertainties derived from 3D position estimation. Additionally, a more accurate target model is also worth to be considered for future investigation. Furthermore, we will record our own large-scale dataset within a challenging outdoor environment to allow for a more realistic evaluation.

Acknowledgements: We would like to thank our colleagues for their comments and suggestions.

References

1. Felzenszwalb, P., Girshick, R., McAllester, D.: Cascade object detection with deformable part models. In: CVPR. (2010)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
3. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Online muti-person tracking-by-detection from a single, uncalibrated camera. PAMI **10** (2010) 1–14
4. Huang, C., Wu, B., Nevatia, R.: Robust object tracking by hierarchical association of detection responses. In: ECCV. (2008)
5. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. International Journal of Computer Vision **75(2)** (2007) 247–266
6. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: CVPR. (2011)
7. Khan, S.M., Shah, M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: ECCV. (2006)
8. Kim, K., Davis, L.S.: Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In: ECCV. (2006)
9. Berclaz, J., Fleuret, F., Fua, P.: Robust people tracking with global trajectory optimizaton. In: CVPR. (2006)
10. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multi-camera people tracking with a probabilistic occupancy map. IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008) 267–282
11. Hartley, R., Zisserman, A.: Multiple View Geometry. the United Kingdom at the University Press, Cambridge (2006)
12. Andriyenko, A., Roth, S., Schindler, K.: An analytical formulation of global occlusion reasoning for multi-target tracking. In: ICCV Workshops. (2011)
13. Nummiaro, K., Koller-Meier, E., Gool, L.V.: Color features for tracking non-rigid object. In: ACTA Automatica Sinica. (2003)
14. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. PAMI **25** (2003) 564–577
15. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The clear mot metrics. EURASIP J. on Image and Video Processing (2008) 10
16. Yang, J., Shi, Z., Vela, P., Teizer, J.: Probabilistic multiple people tracking through complex situations. In: IEEE Workshop Performance Evaluation of Tracking and Surveillance. (2009)