Computer Vision and Image Understanding 134 (2015) 48–63

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Analysis-by-synthesis: Pedestrian tracking with crowd simulation models in a multi-camera video network

Zhixing Jin *, Bir Bhanu

Center for Research in Intelligent Systems, University of California, Riverside, CA 92521, United States

ARTICLE INFO

Article history:
Received 2 February 2014
Accepted 6 October 2014

Keywords:
Pedestrian tracking
Multi-camera systems
Crowd simulation

ABSTRACT

For tracking systems consisting of multiple cameras with overlapping field-of-views, homography-based approaches are widely adopted to significantly reduce occlusions among pedestrians by sharing information among multiple views. However, in these approaches, the usage of information under real-world coordinates is only at a preliminary level. Therefore, in this paper, a multi-camera tracking system with integrated crowd simulation is proposed in order to explore the possibility to make homography information more helpful. Two crowd simulators with different simulation strategies are used to investigate the influence of the simulation strategy on the final tracking performance. The performance is evaluated by multiple object tracking precision and accuracy (MOTP and MOTA) metrics, for all the camera views and the results obtained under real-world coordinates. The experimental results demonstrate that crowd simulators boost the tracking performance significantly, especially for crowded scenes with higher density. In addition, a more realistic simulation strategy helps to further improve the overall tracking result.

1. Introduction

Tracking pedestrians has been an active research topic in computer vision. Although many sophisticated techniques have been proposed, it is still a challenging problem that requires further advances to track people in a camera network in real-world applications. There are many reasons that make tracking a difficult problem. For example, the illumination conditions are changing continuously; the appearances for a single pedestrian are not the same from different perspectives; the amount of occlusions among pedestrians is significant when people form a crowd; and moreover, the human behaviors are sometimes unpredictable. To overcome these problems, researchers have proposed different approaches, among which are the approaches that aim to use multiple overlapping cameras [1]. By fusing information from cameras with overlapped field-of-views, the tracking accuracy could be improved due to the reduction of the influence from occlusion, the separation of crowded people from one another in large foreground blobs, etc. [2,3].

In this paper, we are focused on the setting where several cameras with overlapped field-of-views are used to track pedestrians [4,5]. To take advantage of this setting, many approaches make use of homography-related methods [3,6,7], which are able to model the relationship among different views in order to estimate the actual position of each pedestrian in the real-world ground plane. These methods are efficient since the only extra information that we need to know is the set of camera parameters and the calculation of the perspective transformation is computationally light.

However, in almost all the current trackers that use multiple cameras, the estimated real-world positions for pedestrians are only used in data association across cameras, while their relationships with one another are somehow ignored. In fact, the real-world positions of pedestrians are capable of providing more constraints and predictions, which can be quite helpful in addition to a traditional frame-based tracking approach. From this perspective, crowd simulation is a good example of methods that integrate extra information brought by the real-world positions of pedestrians. In the area of computer graphics, crowd simulation is a very popular topic, with various applications such as designing emergency evacuation routes and introducing special effects in movies. It is used for simulating the behavior of either every individual or the whole group under certain constraints (e.g., to avoid collisions). Nowadays, one of the most popular crowd simulation approaches is mainly focused on simulating walking trajectories (direction and velocity) of each individual given the starting and ending locations. As in a multi-camera system, the direction and velocity information for each pedestrian can be acquired based on the estimated real-world positions at each frame, integrating crowd simulation algorithms with image analysis will be useful for accurate tracking.

In this paper, we propose an approach to improve the performance of a multi-camera tracking system by combining

* Corresponding author.
E-mail addresses: jinz@cs.ucr.edu (Z. Jin), bhanu@cris.ucr.edu (B. Bhanu).

http://dx.doi.org/10.1016/j.cviu.2014.10.001
1077-3142/

vision-based tracking results with the output from a crowd simulation under the real-world coordinates. The system diagram of the proposed approach is illustrated in Fig. 1. In addition to the frame trackers for different cameras which are independent from each other, a crowd simulator runs separately and provides predictions for all the views by projecting the real-world positions back to frame positions. This is a further extension of using the location and velocity information of pedestrians in a tracking system. The frame trackers for each view are designed based on the recent tracking-by-detection approach [8]. Compared to our previous work [9], the manner in which we integrate the information from crowd simulation is different. In addition, the crowd simulator in the proposed approach has two candidates: the RVO2 library [10] and the Social Behavior Model (SBM) [11,12]. Their simulation strategies are different, which may lead to different overall tracking performance. Therefore, besides the utilization of crowd simulation approaches to provide extra predictions for traditional vision-based tracking, a second purpose of this paper is to investigate whether the theoretically better simulation could be more valuable when integrated to the tracking system.

The rest of the paper is organized as follows. Section 2 gives a brief description of related work, including vision based tracking approaches (those use frame information only) and crowd simulation methods, as well as the contributions of this paper. Section 3 presents our proposed approach in detail. Section 4 gives the experimental results and provides a discussion on experiments. Finally Section 5 concludes the paper.

2. Related work and contributions

2.1. Pedestrian tracking

Most of the *state-of-the-art* tracking approaches belong to the category called tracking-by-classification or tracking-by-detection. They usually work in an online manner, e.g., the Online Ada-Boosting [13], Semi-Boosting [14] tracker and Online Multiple Instance Learning [15] tracker. The general idea for this category of trackers is to train a classifier based on the object's appearance features extracted from the initial patch (Region-of-Interest, ROI) on the first frame or the first several frames. Later at each time step, based on the evaluation from this classifier, a patch that maximizes the likelihood in the search window is located, and is further used to update the classifier itself. By repeating these steps, the classifier has the capability to adapt itself to the most recent tracking environment when video continues, as well as maintain a good performance on distinguishing target objects from the surroundings. However, such an evolving classifier may slowly drift and becomes off-target finally. Therefore, some tracking-by-detection approaches perform tracking purely based on detection results from a human detector, which are obtained independently from frame to frame [16–19]. There are also tracking approaches that integrate both evolving classifiers and human detectors [8,20]. Since a human detector is generally more confident, and works independently from the tracker, the input from a human detector is an effective way to initialize the classifier and/or provide the "ground-truth" to correct a drifted tracker. In the proposed approach, the pedestrian tracking method used is based on [8], which is composed of a particle filter, a human detector and an online boosting classifier. This approach is mainly based on particle filtering, while the human detector and online boosting classifier are used to adjust weights for the particles. In addition to the good performance shown in [8], another advantage of this approach is that the crowd simulation can be easily integrated into the system as the extra simulation information can be treated as an additional factor that influences particle weights.

As the ability for tracking in a single-camera is often limited, more cameras are added into a surveillance system to solve the problem, which is known as multi-camera tracking. Basically, there are three settings that a multi-camera tracking system can employ: overlapping cameras [2–5,21], non-overlapping cameras [22,23], and the mixture of them [24]. Related to our research focus in this

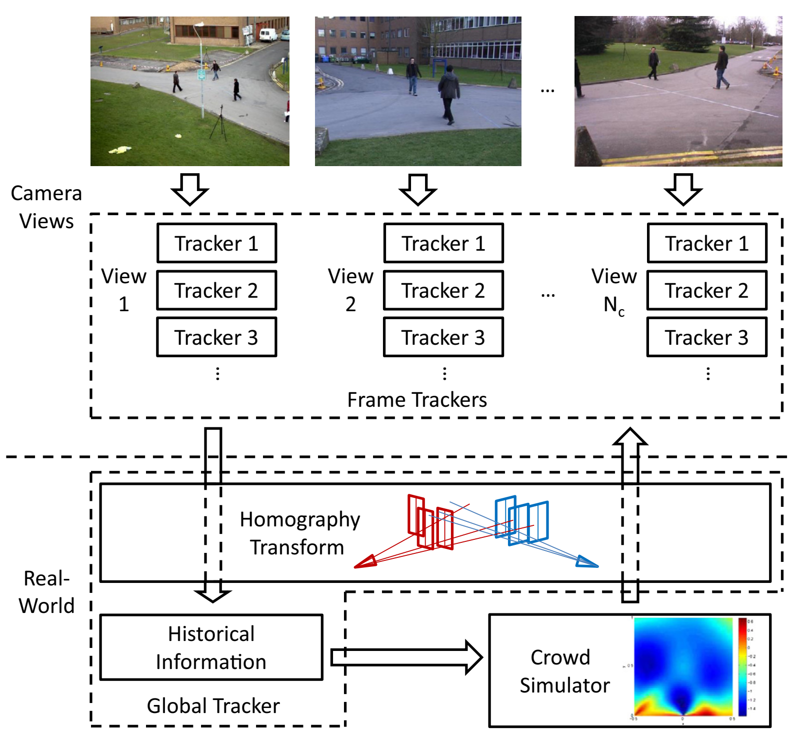

Fig. 1. The system diagram. This example is a system with N_c different views. During the initialization, the information from different camera views is integrated and saved as the historical information. Later at each time step: (1) the crowd simulator first generates the predicted locations for pedestrians; (2) then the simulated locations are passed to different camera views to help update the frame trackers; (3) finally the new position and velocity information for each pedestrian on the ground plane is calculated and used to update the historical information.