Multi-person tracking with overlapping cameras in complex, dynamic environments

Martijn Liem
http://www.science.uva.nl/~mliem
Dariu M. Gavrila
http://www.gavrila.net

Intelligent Systems Laboratory,
Faculty of Science,
University of Amsterdam

The visual tracking of people is an essential capability in many surveillance applications. In this paper, we are interested estimating ground plane location in the more challenging cases involving multiple persons and dynamic environments (e.g. uncontrolled, outdoor settings). To cope with occlusions we use three overlapping, off-line calibrated cameras. Furthermore, an appearance model is used to help disambiguate the assignment of new measurements to existing tracks.

Person tracking has been extensively studied, primarily from a single camera perspective [5, 10]. Previous work has also dealt with tracking persons across multiple cameras and the associated hand-off problem [9]. Regarding the use of overlapping cameras, one of the main methods for determining spatial positions is to geometrically transform images based on a predetermined ground plane homography [1, 3, 6]. Tracking can in these cases be done based on the estimated ground plane positions. Previous work can be distinguished by the type of state estimation employed, whether recursive (Kalman [1, 8], particle filtering) or in batch mode (Viterbi-style MAP estimation [4], graph-cut segmentation in space-time volume [6] or otherwise [2]).

In our approach, position measurements are obtained by carving out the space [7] defined by foreground regions in the overlapping camera views. The 3D space carved by three cameras is subsequently projected onto the ground plane and a connected component analysis finds the associated blobs. Those blobs of a certain width and height, and with sufficient accumulated mass in the vertical direction are considered to represent persons. In practice, a detected blob can be significantly larger than the average size of a single person, when it represents multiple persons, or if it is enlarged by volume carving artefacts. In this case, we use EM clustering to split the blob into multiple sub parts, in such a way that each sub part has the size of an average person.

Since ambiguities can cause non-corresponding foreground regions in different views to be matched, additional volumes will be carved out and projected onto the ground plane. The resulting artefacts have similar characteristics as the blobs corresponding to actual persons, we term them *ghosts*. As a partial solution to reduce the occurrence of *ghosts*, a fixed boundary along which persons can enter the scene is defined. In our case, we define a border area enclosing the space that is visible by all the three cameras. This boundary is not only useful for the detection of *ghosts*, but can also be used for the detections of lost tracks due to people leaving the scene.

The 3D space carved by all three cameras can be projected back onto the respective image planes, obtaining an improved, 'cleaned' binary foreground image. Here, some of the regions associated to background changes are potentially removed. Depth ordering furthermore allows to determine which foreground pixels derive from which 3D objects, and enables the definition of respective occlusion masks. These masks can be used to obtain an accurate appearance measurement for each person in the scene, i.e. only containing data from that individual.

We compute three colour histograms per object, roughly corresponding to typical legs, torso and head/shoulders levels. These act as the person's appearance model. Learned object histograms are compared to the appearances of newly found blobs using the Bhattacharyya distance. Matched appearance models are updated using an exponential decay function.

The assignment of measurements to tracks (modelled by Kalman filters) is done in a non-greedy, global fashion based on ground plane position and colour appearance. The likelihood of the assignment of a tracker to a blob is determined by weighting the euclidean distance between trackers and blobs by their Bhattacharyya distance. A best-first search approach is used to find the optimal assignment in the space of all possible assignments. The advantage of the proposed approach is that the decision on correspondences across cameras is delayed until it can be performed at the object-level in a more robust manner, as compared to matching individual feature locations. This allows us to handle complex environments, containing a significant amount of lighting and background changes, with a moderate number of cameras.

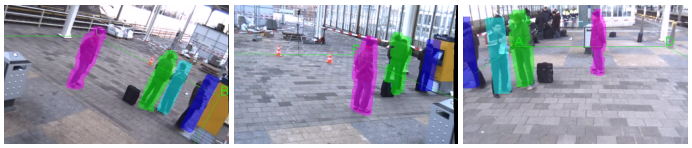

Figure 1: Example results from scenario 9-3. Although one of the people is occluded in the second image, he is still tracked correctly.

The trackers that have not been assigned to blobs, update their state (position) using the Kalman prediction only. Conversely, when after assigning tracks to blobs, additional blobs remain which have not been assigned a tracker, a new tracker is created for that blob and the appearance model is initialized. To reduce the probability that a *ghost* will be tracked, trackers are initialized in a 'hidden' state for the first 20 frames. Since an actual person blob should be well segmented for a longer period of time, ghosts can often be detected by their instability over time.

The setting for our experiments is an actual train station platform on a normal business day. The scenarios recorded for our experiments show two to four actors engaged in different levels of interaction. This ranges from walking by each other, hugging each other to getting pickpocketed. At the same time, a lot of non-scripted activity is going on in the background. Trains and metros are passing by, as well as bystanders who are walking around and getting in and out of trains. Furthermore, lighting conditions change continuously due to the open nature of the location.

An example of a detection result can be seen in figure 1. Overall, our system does reasonably well. In general, the more people there are in the scene, the more ghosting will appear and the higher the risk of false positives. False negatives typically occur because a tracker gets switched onto a ghost object, which also partly explains the rising amount of ID changes when more people come into play. Concluding we can say that on challenging outdoor data involving sizeable changes in lighting and background, we obtained a tracking performance that seems at least on-par with the state of the art. The addition of an appearance model proved to help in the disambiguation of the assignment of measurements to tracks.

[1] D. Arsic et al. Applying multi layer homography for multi camera person tracking. In *ICDSC*, 2008.

[2] S. Calderara et al. Bayesian-competitive consistent labeling for people surveillance. *PAMI*, 30(2):354–360, 2008.

[3] R. Eshel and Y. Moses. Homography based multiple camera detection and tracking of people in a dense crowd. In *CVPR*, pages 1–8, 2008.

[4] F. Fleuret et al. Multicamera people tracking with a probabilistic occupancy map. *PAMI*, 30(2):267–282, 2008.

[5] I. Haritaoglu, D. Harwood, and L. S. Davis. W^4: Real-Time Surveillance of People and Their Activities. *PAMI*, 22(8):809–830, 2000.

[6] S. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint,. In *ECCV*, 2006.

[7] K. Kutulakos and S. M. Seitz. A theory of shape by space carving. *IJCV*, 38(3):199–218, 2000.

[8] A. Mittal and L. Davis. M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. *IJCV*, 51(3):189–293, 2003.

[9] B. Prosser et al. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *BMVC*, 2008.

[10] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *PAMI*, 26(9):1208–1221, 2004.