Multi-person tracking with overlapping cameras in complex, dynamic environments[1]

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

Intelligent Systems Laboratory,
Faculty of Science,
University of Amsterdam

Abstract

This paper presents a multi-camera system to track multiple persons in complex, dynamic environments. Position measurements are obtained by carving out the space defined by foreground regions in the overlapping camera views and projecting these onto blobs on the ground plane. Person appearance is described in terms of the colour histograms in the various camera views of three vertical body regions (head-shoulder, torso, legs). 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 advantage of the proposed approach is that the decision on correspondences across cameras is delayed until it can be performed at the object-level, where it is more robust.

We demonstrate the effectiveness of the proposed approach using data from three cameras overlooking a complex outdoor setting (train platform), containing a significant amount of lighting and background changes.

1 Introduction

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). See the train station setting depicted in Figure 1. Although state of-the-art background modelling algorithms are used, foreground segmentation is clearly affected by sudden lighting changes and non-stationary backgrounds (people moving in the background, trains passing by), and we are interested in a tracking system that is robust to these artefacts. This is achieved by using an appearance model which helps disambiguate the assignment of new measurements to existing tracks. To cope with occlusions we use a moderate number of overlapping cameras (three), which makes it more realistic for surveillance applications, where camera unit cost has come done, but still remains appreciable. The cameras are calibrated off-line.

[1]This research was supported by the Dutch Science Foundation NWO under grant 634.000.432 within the ToKeN2000 program.

BMVC 2009 doi:10.5244/C.23.87

Figure 1: (top row) Three camera views of a train platform involving multiple persons and a dynamic environment (bottom row) Foreground segmentation is clearly affected by sudden lighting changes and non-stationary backgrounds (people moving, trains passing by).

2 Related Work

Person tracking has been extensively studied, primarily from a single camera perspective (e.g. [5, 16]). Previous work has also dealt with tracking persons across multiple cameras and the associated hand-off problem (e.g. [10, 14, 15]). 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 [3, 6, 9]. In this case, the spatial organization of the scene is estimated by projecting segmented 2D objects on the ground plane of the scene using camera calibration and homogeneous transformation. This results in a 2.5D-like approach, where the 2D images are combined into a 2D projection of a 3D scene. Tracking can then be done based on the estimated ground plane positions. In [1] the method of using homogeneous projections of foreground blobs for detecting people is extended by using multiple projection levels. Not only are projections made on the ground plane, but multiple height levels are defined at which images are homogeneously transformed and compared, thus creating a 3D stack of 2D projections giving much more detail on the likelihood of a person's position.

Another, comparable method is proposed by [4]. In this case, stochastic models are used to estimate a ground plane occupancy map, which is used to track people. Camera calibration is needed to find a common ground plane map in all images, but the object segmentations are not transformed onto this ground plane. Instead, the 2D segmentations of the separate camera images are directly used to estimate the occupancy probability of certain prefixed ground plane locations.

Instead of using the ground plane projection to track, it is also possible to track in the 2D camera images and use inter-camera matching information to relate objects in various camera images [2]. In this case, object positions are determined in each individual camera image, tracked separately and then matched between camera's, based on appearance model and geometrical features. This kind of method is especially useful for consistent labelling of people over multiple cameras and a long period of time.

Another possibility is to use the reconstructed 3D scene for tracking, occlusion detection and person identification. A method using epipolar geometry to construct the 3D scene is proposed by [13]. In this case, colour similarities of segmented objects along corresponding epipolar lines are used to identify people positions in 3D space ([14] proposes a more sophisticated colour matching across cameras by means of cumulative brightness transfer function). A top down projection of the scene is used for tracking people.

Overlapping camera systems were also applied for 3D pose recovery in controlled indoor environment. The near-perfect foreground segmentation resulting from the stationary (blue screen type) background, together with the many cameras used (> 5), allowed to recover pose by Shape-from-Silhouette techniques [8, 17].

Previous work can be furthermore distinguished by the type of state estimation employed, whether recursive (Kalman [1, 7, 13], particle filtering) or in batch mode (Viterbi-style MAP estimation [4], graph-cut segmentation in space-time volume [9] or otherwise [2]).

In our approach, position measurements are obtained by carving out the space [11] defined by foreground regions [17] in the overlapping camera views. Person appearance is described in terms of the colour histograms in the various camera views of three vertical body regions (head-shoulder, torso, legs). 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 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 (i.e. matching entire object appearance at once, and not requiring colour correspondences across cameras), as compared to matching individual epipolar line segments [13] or individual height layers [1, 5, 9]. This allows us to handle complex environments, containing a significant amount of lighting and background changes, with a moderate number of cameras, see Figure 1.

3 Person Tracking based on Position and Appearance

3.1 Person Position

Position measurements are obtained by carving out the 3D space defined by foreground regions in the overlapping camera views. See Figure 2 (left) for a schematic overview. 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 (user-set thresholds), are considered to represent persons. An example of such top-down view is seen in Figure 2 (right).

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 split the blob into multiple sub parts, in such a way that each sub part has the size of an average person, and leave the interpretation up to the tracker in its assignment of measurements to active tracks, see Section 3.3. The blob-splitting is done by Expectation Maximization (EM) clustering on the blob. The number of Gaussian kernels to be used for EM are estimated by comparing the blob size to the minimum size of a one-person blob. Currently, we use four times the minimum blob size as the threshold for adding an extra kernel. After having split all large blobs, we have a list of possible locations of individual people.

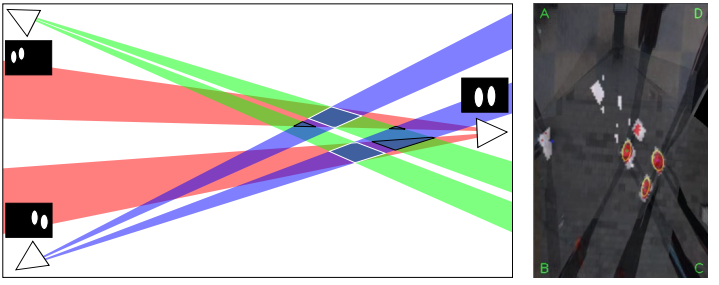

Figure 2: **(left)** Space volume carving projects foreground masks generated at each camera view into a 3D space, 'carving out' the positions where objects might be (white bounds). Because of uncertainty in the carving result, extra areas will get carved out which will appear as 'ghosting' artefacts (black bounds). **(right)** Ghosting due to background noise and the scene layout. While this scene contains only three actual people, multiple red blobs are visible in the top-down view. Furthermore, one of the three people is actually at the centre left edge of the screen, while no blob is visible in that area due to segmentation errors. The white areas in the top-down view are artefacts created by noise and people segmentations erroneously matching from different camera viewpoints, which have been filtered out because they lack the mass necessary to represent a person. This image gives an example of why the segmentation of people based on a top-down projection from the 3D reconstructed scene can be hard at times.

Due to incorrect correspondences between the foreground images across camera views (involving multiple persons and/or erroneous foreground regions), 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*. See Figure 2. *Ghosts* are often gradually created and merged back into real person blobs, making it hard to distinguish them from groups of people getting close together and parting ways. Thus, the detection of *ghosts* on a temporal basis of position is difficult.

A partial solution to reduce the occurrence of *ghosts* is to define a fixed area where persons can enter the scene. In our case, we define a border area enclosing the space that is visible by all the three cameras. The sudden appearance of new persons in the middle of the scene can be disregarded, preventing the system from tracking incorrect objects. Another, more general solution is the use of an appearance model to distinguish between real persons and *ghosts*, see next Subsection.

3.2 Person Appearance

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 (cf. train passing by in Figure 1 (middle view)). 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 colour histograms ($10 \times 10 \times 10$ bins RGB) of the projection of the 3D objects in the three camera views, at three height intervals above the ground plane, which

roughly correspond to typical legs, torso and head/shoulders proportions. The histograms from the separate cameras are merged into one (normalized), which acts as the person's appearance model. The combined histogram representing a possible person location and the one used to describe the person appearance at a previous time for a certain track are compared by the Bhattacharyya distance. If the new measurement is assigned to the existing track, the appearance model is updated by an exponential decay function.

3.3 Person Tracking

Detected blobs that lie within a certain distance to track predictions are candidates for assignment to the corresponding trackers (gating). Using the positions of the trackers and the blobs, as well as the similarity between the trackers' known appearance models and the blobs' appearances measured using the Bhattacharyya distance D_b, the best match between blobs and trackers is determined. For this, the Euclidean distance D_e between each blob and the predicted new position of each tracker's Kalman filter is weighted using the Bhattacharyya distance between them, which has a value between 0 and 1. The resulting value can be used to determine the likelihood of the assignment of a tracker to a blob. We use best-first search to search in the space of all possible assignments. The likelihood value of a particular (partial) assignments is used to determine which assignment is evaluated next by the best-first search algorithm.

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 (see Section 3.2). 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. Taking this into account, we only make a tracker 'visible' when it has been able to track a blob 20 frames in a row. Within this timespan, the blob is taken into account when computing appearance models. When the track is lost before its 20^{th} frame existence, it is discarded.

As mentioned earlier, new tracks are only created at the border of the detection area (i.e. area visible by all three cameras). This helps us prevent the assignment of tracks to *ghosts*, appearing in the middle of the scene. Furthermore, when a tracker is lost outside the detection area for more than 20 frames, the person is assumed to have disappeared from the scene and the track is deleted. Notice that we are not able to re-identify people leaving the scene and re-entering in a completely different position.

Our tracking system is summarized by Algorithm 1.

4 Experiments

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 train platform. In total, 8139 frames in 6 scenarios were recorded and analysed.

Algorithm 1: The tracking algorithm

Input: Three calibrated RGB camera images overlooking a particular scene
Output: Positions of persons in the scene

foreach *time step t* **do**
- Do background estimation on all three images to get foreground regions;
- Execute volume space carving to get a 3D scene representation;
- Project 3D space onto the ground plane;
- **foreach** *blob b > minimum person size* **do**
  - Do EM using as many kernels as there are possibly persons in the blob;
  - Replace b by its sub-blobs;
  - Compute appearance model for each (sub-) blob;
- Compute euclidean distance $D_{e_{k,b}}$ between all b and k;
- Compute Bhattacharyya distance $D_{b_{k,b}}$ between appearances A_b^t of all b and A_k^{t-1} of all k;
- **forall** *combinations of k and b where* $D_{e_{k,b}} < 1.2m$ *AND* $D_{b_{k,b}} < 0.4$ **do**
  - Find optimum configuration of unique assignments using best-first one-to-one matching of k and b using $D_b D_e$ as the distance measure;
- **forall** *assigned trackers k* **do**
  - Update Kalman filter using blob position;
  - Update appearance model $A_k^t = (1-\alpha)A_k^{t-1} + \alpha A_b^t$;
- **forall** *unassigned trackers k* **do**
  - **if** *k exists < 20 frames OR k behind entry bounds > 20 frames* **then**
    - Discard k;
  - **else**
    - Update k using predicted Kalman filter location;
- **forall** *blobs b without a tracker* **do**
  - **if** *b within entry bounds* **then**
    - Create new tracker k;
    - Set appearance model A_k^t to A_b^t;

scenario	GT	TP	PTP	DR	IDC	FN	FP
1-1	3	3	1	98.5	0	0	0
7-1	4	3	0	87.5	2	1	1
8-1	2	2	0	99.3	0	0	0
9-3	4	3	1	85.5	4	1	2
10-2	3	2	0	85.4	6	1	2
11-1	2	2	0	97.2	0	0	0
Total	18	15	2	92.2	12	3	5

Table 1: Tracking results on our scenes. GT: ground truth (number of persons); TP: true positive rating (person > 75% tracked); PTP: perfect true positive rating (person 100% tracked); DR: detection rate; IDC: ID changes; FN: False negatives (person < 75% tracked); FP: false positives (tracker > 75% without person)

Ground truth was created by manually defining the 3D locations of the persons in the scenes, using the images taken by the three cameras.

The metrics used to assess the performance of our tracker are similar to those used in [3]. We assume a tracker to be assigned to a person as long as the tracker is within 0.75 meters of a person. When a person is tracked for more than 75% of the time they are in the scene, the track is considered to be a True Positive (TP). If the track is correct for 100% of the time, not regarding identity changes, it is even classified as a Perfect True Positive (PTP). A person who is tracked less than 75% of the time however, is considered to be a False Negative (FN), while a track which is not matched to a person for at least 75% of the time is classified as being False Positive (FP). Finally, we count the number of times a different tracker is assigned to a person (ID change, IDC) and the total fraction of the time persons are being detected and tracked (Detection Rate, DR).

Table 1 summarizes the tracking results for the different scenarios. Various screen shots can be seen in figure 3. In scenario 1-1 (869 frames), one person enters the scene and keeps standing still at a fixed position, while two more people enter and hug each other. Although the hugging persons pose a difficult situation because they tend to be seen as one person in the reconstruction, our system is, thanks to the blob splitting policy, able to track all people quite well, of which one (the one standing still) even perfect. This last point seems to be obvious, but due to adaptive nature of the background estimation algorithm [17], people standing still for a longer period of time dissolve into the background. This makes it hard to keep track of this person, since a major clue, the top-down projection of the 3D segmentation, no longer exists. Because we stop updating tracker locations as soon as no input is received and the prediction of the Kalman filter estimates movement to be less than 0.05 meters (with a frame rate of 20 fps, about 1 m/s), we can make sure the tracker stays around the correct location until the person starts moving again.

Scenario 7-1 (2079 frames) shows multiple persons waiting in line for a ticket machine. The long period of time they are standing still causes some confusion in tracker assignment. A person lingering around in the boundary region of the scene causes a false negative, since the tracker constantly gets out of the scenario bounds and is therefore removed. Because multiple people standing still near the same position, foreground segmentation errors cause a tracker to be left behind after a person starts moving again. While the person gets assigned a new tracker, causing an ID change, the tracker left behind creates a false positive.

Scenario 8-1 (517 frames) is a pickpocketing scenario, showing one person standing still at a vending machine when he is pickpocketed by a person passing by. While the people

Figure 3: Screen shots from three camera views showing tracking results (each track shown in same colour across views). From top to bottom: scenario 1-1, 7-1, 8-1, 9-3, 10-2 and 11-1. Note strong lighting changes between first two rows.

get very close together, tracking tends to be near perfect. Since the two persons are tracked for 99.6% of the time and 99.74% of the time resp., they are just a tiny bit away from being classified as a perfect true positive. 100% is not reached because one of the people walks just past the boundary of the scene, losing the tracker for a moment after which the tracker is relocated on the person.

In scenario 9-3 (1188 frames), multiple persons standing still as well as several ghosts occurring because of people standing close together cause some incorrect detections and switching of tracks between persons. While overall the tracking is not bad, even having a perfect true positive detection, the large amount of activity in the same neighbourhood as where people are standing still cause some trackers to run off and cause tracking errors.

Scenario 10-2 (2424 frames) is one of the more difficult scenarios. Difficulties with ghosting and people standing still for long periods of time cause a lot of switching identities. The main challenge is in correctly assigning tracks of people who are no longer visible as a blob, because they have been standing still for too long, while a ghost pops up just near the last known location of the person. Preventing the tracker from switching over turns out to be quite hard in these cases.

Finally, scenario 11-1 (1062 frames) shows a near-perfect tracking result. There are two people in the scene, arguing and moving around near the same spot. Only at the very last moment, one of the tracker fails to pick up on one of the persons who faded away a bit into the background because of the time spent near the same position.

Overall, our system does reasonably well. Compared to the results presented by [3] when using only three cameras, our results seem comparable and with respect to some points even better (e.g. [3] list in their Figure 6 a detection rate DR of 73%). However, direct comparisons are difficult since the data sets are different.

There is a clear relation between the length of a sequence, the number of persons appearing in the sequence and the quality of the tracks. 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. While the appearance model can prevent some of those switches from happening, it is hard to identify those ghosts which were created by mismatching two similarly dressed persons in the scene.

5 Conclusion

We presented a system for tracking multiple people using three calibrated cameras. Volume carving and projection onto the ground plane determined potential people positions; these were used as input to tracking. 3D scene reconstruction and depth-ordering provided the respective occlusion masks for measuring person appearance. Our person appearance model described three vertical body regions (head-shoulder, torso, legs) in terms of the colour histograms in the various camera views.

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. In order to further reduce ID switches, we plan to investigate stronger appearance models that more strongly link the spatial and texture cues.

References

[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, Augustus 2000.

[6] W. Hu et al. Principal axis-based correspondence between multiple cameras for people tracking. *PAMI*, 28(4):663–671, 2006.

[7] J. Kang, I. Cohen, and G. Medioni. Tracking people in crowded scenes across multiple cameras. In *ACCV*, 2004.

[8] R. Kehl and L. Van Gool. Markerless tracking of complex human motions from multiple views. *CVIU*, 103(2-3):190–209, 2006.

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

[10] S. Khan et al. Human tracking in multiple cameras. In *ICCV*, 2001.

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

[12] I. Mikic et al. Human body model acquisition and tracking using voxel data. *IJCV*, 53 (3), 2003.

[13] 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.

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

[15] W. Zajdel, Z. Zivkovic, and B. J. A. Kröse. Keeping Track of Humans: Have I Seen This Person Before? *ICRA*, pages 2081–2086, April 2005.

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

[17] Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 27(7):773–780, May 2006.