Multi-Person Localization and Track Assignment in Overlapping Camera Views

M. Liem[1] and D. M. Gavrila[1,2]

[1]Intelligent Systems Lab, Fac. of Science, Univ. of Amsterdam, The Netherlands
[2]Environment Perception, Group Research, Daimler AG, Ulm, Germany

Abstract. The assignment of multiple person tracks to a set of candidate person locations in overlapping camera views is potentially computationaly intractable, as observables might depend upon visibility order, and thus upon the decision which of the candidate locations represent actual persons and which do not. In this paper, we present an approximate assignment method which consists of two stages. In a hypothesis generation stage, the similarity between track and measurement is based on a subset of observables (appearance, motion) that is independent of the classification of candidate locations. This allows the computation of the K-best assignment in low polynomial time by standard graph matching methods. In a subsequent hypothesis verification stage, the known person positions associated with the K-best solutions are used to define the full set of observables, which are used to compute the maximum likelihood assignment. We demonstrate that our method outperforms the state-of-the-art on a complex outdoor dataset.

1 Introduction

We are interested in tracking a handful of persons in dynamic, uncontrolled environments using overlapping cameras[1]. Cost and logistics typically limit the number of cameras that can be used, as well as their viewpoints. We aim for methods that can cope with as few as three surrounding cameras and diagonal viewing directions that maximize overlap area (as opposed to ceiling-mounted cameras with a bird-eye's view). The considered set-up makes it difficult to establish individual feature correspondences across camera views, furthermore, inter-person occlusion can be considerable. We aim for robustness by performing detection and tracking based on a 3D scene reconstruction, obtained by volume carving [14]. A main challenge is to establish correct object correspondence across multiple views. Matching different objects together across multiple views leads to erroneous 3D objects, so-called 'ghosts' (see Figure 1).

2 Previous Work

Person tracking has been studied extensively. Due to space limitations, we restrict ourselves to work using overlapping cameras that aims to recover multi-

[1] This research received funding under EC's FP7/2007-2013 under grant agreement nr. 218197 (ADABTS).

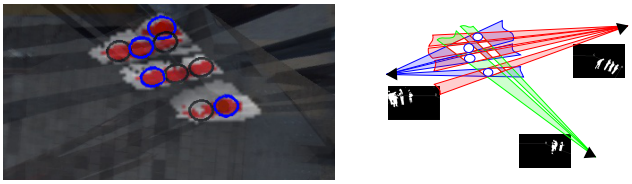

Fig. 1. (left: real-world, right: schematic) Volume carving [14] projects foregrounds for all cameras into a 3D space, 'carving out' potential persons (left: red areas, right: red bounded, white areas). Splitting these into individual potential person measurements results in superfluous objects caused by incorrect correspondences ('ghosts' or artifacts, black ellipses (left), unmarked white areas (right)) and actual persons (blue ellipses).

person location. See Table 1 for an overview that highlights the way person localization and track assignment is performed - our primary paper scope. These approaches can thereafter be embedded in a state estimation framework, either recursive (Kalman [1][8][12], particle filtering) or in batch mode (Viterbi-style MAP estimation [6], graph-cut space-time segmentation [9] or otherwise [4, 10]).

Apart from the various ways correspondence and localization is performed the main point to note from Table 1 is that multi-person localization and track assignment is performed in a decoupled manner. This means that person localization does not take advantage of motion and appearance cues associated with active tracks; only *after* person position has been determined are the latter cues incorporated for track assignment [6]. This approach faces difficulties in disambiguating tracks in close proximity. Therefore, in this paper, we pursue person localization and track assignment jointly. A similar concept was proposed in [10] in a single view context. However, only pairwise object interactions were taken into account while leaving out the dependency between the perceived object appearance and the selected hypotheses. Here, we consider an instantiation specific to multi-view tracking, and we propose a novel two-stage joint estimation procedure to handle the potentially unfavorable (exponential) complexity.

3 Multi-Person Track Assignment

To treat person localization and track assignment jointly, we formulate the problem as an edge selection task on a bipartite graph $G = (X, Z, E)$ with vertex sets X and Z and edges E. Given m measurements of potential persons (see figure 1), n currently existing tracks, p possible track creations and r possible track terminations, each set contains $v = \max(n, m) + \max(p, r) + 1$ vertices. Vertex set $X = \{x_1 \dots x_n, \pi_1 \dots \pi_p, \gamma_1 \dots \gamma_{v-n-p}\}$ contains vertices x_i for existing person tracks, π_i for the generation of new person tracks, and γ_i for the generation of a false positives ('ghosts'). Vertex set $Z = \{z_1, \dots, z_m, \omega_1, \dots \omega_r, \delta_1 \dots \delta_{v-m-r}\}$ contains vertices z_j for measurements, ω_j corresponding to terminated tracks, and δ_j to represent erroneous (i.e. noise) measurements. The bipartite graph has edges E such that: (1) all edges $e \in E$ connect vertices from X and Z: $e \in X \times Z$, (2) vertices within X and Z have degree one (i.e. are connected by one edge) and (3) E does not contain edges connecting a vertex π_i to ω_j.

Method	CA	NP	Localization	Track Assignment
Arsić [1]	4	5	foreground segmentation, multi-plane homography, basic false pos. reduction	quadratic programming: position + appearance (SIFT)
Berclaz [2]	4	5	person classifier, prob. occupancy map	-
Calderara [4]	4	3	homography, epipolar constraints, appearance based	-
Eshel [5]	9	21	homography, intensity corr., false pos. reduction only during tracking	position + appearance
Fleuret [6]	4	4-6	foreground segmentation, probabilistic occupancy map	foreground segmentation + position + appearance (color hist.)
Hu [7]	2-3	4	foreground segmentation, principal axis, homography	position
Kang [8]	2	5	foreground segmentation, homography	multi-hypothesis (JPDA): 2D and 3D position + appearance (color descr.)
Khan [9]	4	9	foreground likelihood homography	space-time segmentation
Liem [11]	3	4	foreground segmentation, volume carving, no false positive reduction	nearest neighbor: position + appearance (color hist.)
Mittal [12]	4-16	3-6	color matching of epipole segments	position, velocity
Yang [15]	8	8	foreground segmentation volume carving basic false positive reduction	-
This method	3	2-4	joint person localization and assignment: foreground segm., volume carving + appearance (color hist.) + position Hungarian method and combinatoric approach	

Table 1. Overview of multi-person localization and track assignment using overlapping cameras (CA: Number of cameras, NP: Number of persons).

The set E can be divided into subsets E^C, E^N, E^D, and E^G, containing

- $\langle x_i, z_j \rangle \in E^C$: z_j is the person assigned to continued track x_i,
- $\langle \pi_i, z_j \rangle \in E^N$: z_j is a person which should be assigned to a new track,
- $\langle x_i, \omega_j \rangle \in E^D$: track x_i can be deleted,
- all other edges $\in E^G$: involving 'ghosts'.

Furthermore, we set $p = r = 1$, thus allowing the addition/removal of only one person track per frame. At a framerate of 20 Hz, this means that 20 persons could be added or removed every second. We also ensure that X and Z have at least one vertex γ_i and δ_j by setting $v = max(n, m) + 2$ in our experiments.

3.1 Likelihood formulation

A set of features O is derived from the measurements. This set consists of the foreground image regions O^{FG}, the position on the ground plane O^{Pos} and appearance O^{App} of (possible) persons. For a given set of edges in the bipartite graph, we model the probability of observing these features:

$$p(O|E) \;=\; p(O^{Pos}|E)\; p(O^{FG}|E)\; p(O^{App}|E). \tag{1}$$

The probability distribution over the positions of measurements only depends on the position of the assigned tracks, or the position where a new track is created or removed.

$$p(O^{Pos}|E) = \prod_{e_k \in E^C} p(O_k^{Pos,C}|e_k) \times \prod_{e_k \in E^N} p(O_k^{Pos,N}|e_k) \times \prod_{e_k \in E^D} p(O_k^{Pos,D}|e_k) \times p_{nPos}^{|E^G|} \tag{2}$$

$O_k^{Pos,C}$ denotes the deviation between predicted location of a track and the position of a measurement on the ground plane. $O_k^{Pos,N}$ denotes the measured position of a new track on the ground plane. $O_k^{Pos,D}$ denotes the disappearance of a measurement. p_{nPos} is a penalty factor, given by the likelihood at the particular distance where $p(O^{Pos}|E^G) = p(O^{Pos}|E^C)$. Note that $p(O_k^{Pos,C}|E) = p(O_k^{Pos,C}|e_k)$ and $p(O_k^{Pos,N}|E) = p(O_k^{Pos,N}|e_k)$ (i.e. $p(O_k^{Pos,N}|e_k)$ does not depend on any $e \in E \setminus e_k$).

We expect that tracked persons explain the observed foreground regions O^{FG} in each camera view. Following [6], the foreground observation probability in a camera c is $p(O_c^{FG}|E) = \frac{1}{Z}e^{-\Psi(B_c,A_c(E))}$, where $A_c(E)$ denotes the synthetic image obtained by putting rectangles at locations corresponding to z_j for which $e_k \in E^C \cup E^N$ (i.e. the union of the corresponding rectangles), B_c is the segmented foreground region, and $\Psi(B_c, A_c(E))$ the fraction of the foreground correctly segmented (c.f. [6]). Averaging over all C cameras results in

$$p(O^{FG}|E) = \frac{1}{C}\sum_{c=1}^{C} p(O_c^{FG}|E) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{Z}e^{-\Psi(B_c,A_c(E))}. \tag{3}$$

If all z_j are outside the field of view of camera c, $p(O^{FG}|E)$ is not computable (since $\Psi(B_c, A_c(E))$ contains a division by $|A_c(E)|$). For these cases, a good value for $\frac{A_c(E)\oplus B_c}{A_c(E)}$ in $\Psi(B_c, A_c(E))$, with $\oplus$ the per-pixel exclusive or, was experimentally found to be 1.5. This value is also used for computing the penalty term p_{nFG}, used when the foreground likelihood is not computable (e.g. for E^D).

Appearances are represented as three RGB color histograms ($10 \times 10 \times 10$ bins): for the legs, arms/torso and head/shoulders, respectively. Splitting the appearance vertically allows us to use and update appearance features, even if a person is partially occluded. Spatial occlusion information, based on detected persons in E, is taken into account when sampling the images and updating the tracked appearance. Histograms are taken from each camera viewpoint and averaged over the different viewpoints:

$$p(O^{App}|E) = \prod_{e_k \in E^C}\left[p(O_k^{App}|E)\right] \times p_{nApp}^{|E\setminus E^C|}$$

with

$$p(O_k^{App}|E) = \frac{1}{C}\sum_{c=1}^{C} p(O_{k,c}^{App}|E). \tag{4}$$

where $O_{k,c}^{App}$ is the Hellinger distance [3] (equal to $\sqrt{1-BC}$, with BC being the Bhattacharyya Coefficient) between the appearance of measurement k in camera c and the known appearances of the tracks in E^C. The factor p_{nApp} compensates for the lack of appearance information for objects not linked to existing tracks, represented by the point where $p(O^{App}|E^G) = p(O^{App}|E^C)$. Distributions $O_{ij}^{Pos,C}|e_k$ and $O_{ij}^{App}|e_k$, are determined experimentally on a separate validation set. Values for these distributions are aggregated across C different camera views. See also section 4.1.

3.2 Likelihood optimization

A brute-force approach to finding the most likely set of edges E for (1) would quickly become intractable due to the combinatorial nature of the assignment problem, especially when there are many measurements. Instead, the idea is to only compute the full likelihood on K preselected probable solutions, after which the most likely one is selected as our final estimate. Preselection is achieved by approximating $p(O|E)$ as a function $\hat{p}(O|E)$ that can be written as a product of independent edge likelihoods. An extended version of the Hungarian algorithm [13] finds the top K most likely solutions for $\hat{p}(O|E)$ in the bipartite graph by expressing it as a max-sum problem which can be solved in low polynomial time.

Since (3) and (4) contain terms dependent on the complete assignment E (e.g. due to occlusion), the conditional probabilities $p(O^{App}_{k,c}|E)$ and $p(O^{FG}_k|E)$ are replaced by approximations $\hat{p}(O^{App}_{k,c}|E)$ and $\hat{p}(O^{FG}_k|E)$ respectively where the likelihood of each edge is independent of the other edges.

Instead of taking possible occlusion of people into account, as was the case in (3), $\hat{p}(O^{FG}|E)$ approximates the foreground probability by computing it independently for assigned tracks:

$$\hat{p}(O^{FG}|E) = \prod_{e_k \in E^{cont,new}} \hat{p}(O^{FG}_k|e_k) \times p_{nFG}^{|E \setminus \{E^C \cup E^N\}|}$$

with

$$\hat{p}(O^{FG}_k|e_k) = \frac{1}{C}\sum_{c=1}^{C} \frac{1}{Z} e^{-\Psi(B_c, A_c(e_k))}. \tag{5}$$

Approximation $\hat{p}(O^{App}_{k,c}|E)$ only includes the appearance of measurements k in those camera views $\mathbf{C}^k$ where the appearances are *guaranteed* not to be occluded, such that dependency on E can be dropped:

$$\hat{p}(O^{App}_{k,c}|E) = \begin{cases} p(O^{App}_{k,c}|e_k) & \text{iff } c \text{ in } \mathbf{C}^k \\ p_{nApp} & \text{otherwise} \end{cases} \tag{6}$$

Now (1) is approximated as:

$$\begin{aligned} \hat{p}(O|E) = &\prod_{e_k \in E^C} p(O^{Pos,C}_k|e_k)\hat{p}(O^{FG}_k|e_k)\hat{p}(O^{App}_k|e_k) \times \prod_{e_k \in E^N} p(O^{Pos,N}_k|e_k)\hat{p}(O^{FG}_k|e_k)\ \ p_{nApp} \\ &\times \prod_{e_k \in E^D} p(O^{Pos,D}_k|e_k)\ \ p_{nFG}\ \ p_{nApp} \times \prod_{e_k \in E^G} p_{nPos}\ \ p_{nFG}\ \ p_{nApp}, \end{aligned} \tag{7}$$

which contains a term for each edge independent of the other edges. Using this expression we preselect the K solutions with the Hungarian method.

4 Experiments

4.1 Setup

Experiments were performed in a complex, outdoor setting. On a train station platform, 2 to 4 actors engaged in various activities. The background is dy-

namic (trains are passing by, bystanders are walking around) and lighting conditions change continuously. Ten sequences were used, with about 5300 multi-view frames (avg. distance between center points of closest persons is 1.6 m, std. dev. is 1.2 m). For the purpose of evaluation, we only considered the area visible in all three cameras, see Figure 2(left). Ground truth (torso position) was created by manual labeling.

Proposed Method Space volume carving is used to 'reconstruct' a 3D representation of the objects in the scene, making use of foreground segmented images. All objects are projected onto the ground plane where only those having sufficient vertical mass to represent a person are kept. An object is detected as a possible person when the area of its top-down projection has at least half the size of an average person. Preliminary tests on our data have shown that on average a person has a top-down silhouette approximated by the area of a circle with a 40 cm diameter. The number of possible persons within one object is determined to be the number of times this 'average person' fits into the detected object. The EM algorithm is used to find the most likely positions of multiple persons in objects larger than one person. It is adapted in such a way that it fits an equally sized ellipse for each person, each ellipse having an aspect ratio of 2:3 representing the average human shape seen from top-down.

Parameterizing the likelihood $p(O_k^{Pos,C}|e_k)$ is done by an exponential distribution using $\lambda = 1/0.03$ (estimated by measuring distances between people in a validation set). The steep descent of such a distribution makes high values unlikely, which de facto puts a bound on the distance a person can travel between 2 frames (0.05 seconds). Approximating the distance distribution of non-person objects $p(O^{Pos}|E^G)$ is optimal using a log-normal distribution $\ln\mathcal{N}(0.22, 0.05)$. The largest allowable distance between two objects, still being classified as persons is set at the distance where $p(O^{Pos,C}|E^C) = p(O^{Pos}|E^G)$, which is 0.2 m. This results in a maximum movement speed of about 14 km/h.

Distribution $p(O^{App}|E^C)$ and $p(O^{App}|E^G)$ are described as log-normal distributions having parameter settings $\ln\mathcal{N}(-2.2, 0.6)$ and $\ln\mathcal{N}(-1.0, 0.5)$ respectively. Since the Hellinger distance takes on values between 0 (complete match) and 1 (no match at all), the range of these functions is limited. For the penalty term $p_{nApp}^{|E\setminus E^C|}$, a Hellinger distance of 0.3 is used, representing a likelihood of 0.5.

Finally, $p(O_k^{Pos,N}|e_k)$ and $p(O_k^{Pos,D}|e_k)$ are defined using an inverted distance map (figure 2, right) based on the boundaries of the scene's visible area (figure 2, left). This map assigns high likelihood to person creations and deletions at the borders of the scene and decreases the likelihood according to the distance from the nearest edge. For the penalty term p_{nPos}, a value of 10^{-4} is found to be reasonable.

Comparison Method We compare our proposed algorithm with the Probability Occupancy Map (POM) algorithm, a state-of-the-art method for which the software was kindly made available by the authors of [6]. This system uses the foreground segmented images as returned by our system as input. For each item on a predefined list of discretized ground plane positions, the POM algorithm

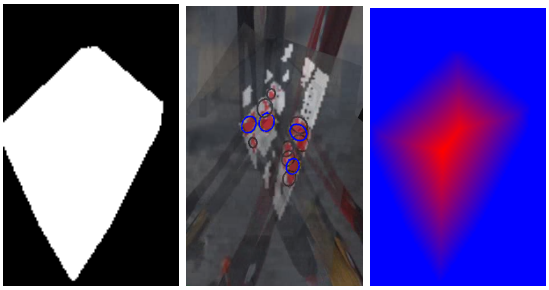

Fig. 2. (left) Area of interest on the ground plane, covered by 3 cameras and used for tracking and detection. (middle) Scene top-down view (similar to figure 1, left) (right) Distance map for determining addition/removal likelihood. Blue: high likelihood, red: low likelihood

returns the likelihood that a person is present at that location. In [6] the ground plane was discretized using a regular grid of size 20 cm. We increased the resolution to 10 cm to compensate for binning effects; this improved performance, especially at low positional error tolerance. Computing the person presence likelihood is done based on the amount of segmented foreground inside a fixed-size Region of Interest (ROI), positioned on each ground plane location. These ROI are represented by boxes of 2 m high and 70 cm wide, projected in each camera. These proportions roughly correspond to those provided in the software by [6] and have been verified to work well in preliminary experiments.

Due to the large grid (9100 locations) and the large number of detections in the neighborhood of a person at the selected likelihood threshold (see next section), computing a match between all persons at t and all detections at $t+1$ would be very costly. In order to keep things manageable, Non-Maxima Suppression (NMS) is used to keep only the most likely person positions in a 3×3 grid neighborhood. Matching is done by evaluating $p(O^{Pos}|E)p(O^{App}|E)$ for all combinations of accepted detections at t and $t+1$. The term $p(O^{FG}|E)$ from (1) is left out of this equation since it is already embedded in the initial POM results [6].

4.2 Evaluation

Detections Both the proposed and POM method have a main parameter that controls the number of candidate person locations that are detected. For our method, this is the minimum vertical mass, for the POM method, this is a threshold on person likelihood at a grid position. In order to find comparable values for the later evaluation of track assignment and tracking, we computed their effect on the True Positives (TP), False Positives (FP) and False Negatives (FN). See figure 3. Based on this, we selected a minimum vertical mass threshold of 90 cm for our method, and a likelihood threshold of 0.01 for the POM method.

Preselection The quality of the proposed preselection (Section 3.2) is tested on a separate validation set (around 10^4 frames, eight scenarios). A cumulative

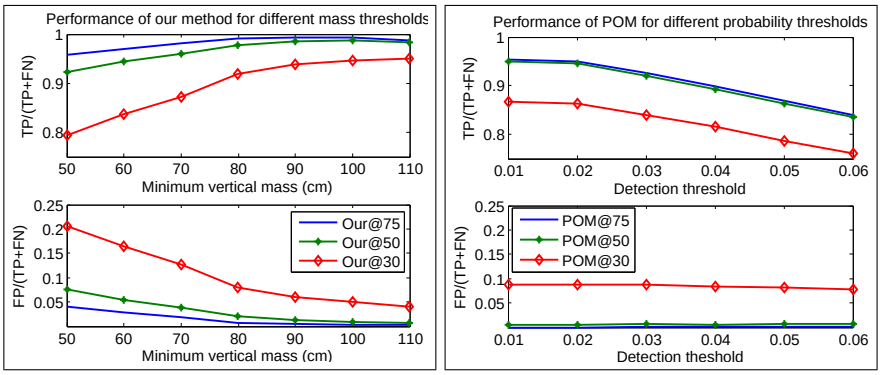

Fig. 3. (left) Our detection performance for different minimum allowed vertical mass (right) POM detection performance for different detection thresholds. Lower threshold values for POM could not be tested since the resulting increase in the number of detections causes computational issues. Both figures show multiple maximum allowable GT to detection distances (30 cm, 50 cm and 75 cm).

plot of the fraction of frames where the correct solution occurs within the first x solutions is given in figure 4 (left). A solution is deemed correct when all Ground Truth (GT) persons are localized in the scene with a maximum distance error of 75 cm and there are no false positives. The results were computed incrementally, i.e. persons detected at time t are based on the result found at $t-1$, which in turn depended on result at $t-2$, etc. (no filtering is performed). The cases where the correct solution was not present among the top-100 ranked solutions are mostly caused by errors in foreground segmentation (this does not necessarily mean that the system loses track from that point on; a tracker might still recuperate). From these experiments, 40 is determined to be a good cut-off point for the number of hypotheses maintained after preselection.

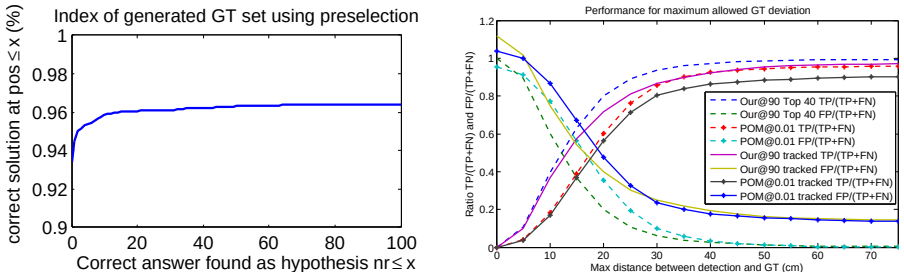

Fig. 4. (left) Percentage of cases that the correct solution is among the top x of solutions produced. (right) Detection and tracking performance of our method and the POM method, given a maximum allowable error between the GT positions and the detected positions.

Person Localization and Track Assignment Performance evaluation is done for both methods on a frame-to-frame basis, i.e. new detections at $t+1$ are

matched to GT person positions from t. This allows us to focus on the person localization and track assignment capability. For a fair comparison, POM detections are used as the input of our system, replacing volume carving. Cylinders of 70 cm diameter and 2 m high (equal to the ROI used by POM, section 4.1) are generated in the voxelspace at the locations of POM's detections. Person detection is done using our two-stage estimation process on this POM-generated voxelspace. Figure 4 (right, dotted lines) shows the performance of our method as well as the POM method, given different maximum allowable errors between the GT and the detections. Our method outperforms POM for any of the tested maximum GT error distances (higher TP, lower FP). This is especially the case for positional tolerances below 30 cm, where the grid-based nature of POM leads to binning artifacts. Even at the highest allowable GT error of 75 cm, our method still has a TP rate about 4% above POM. This is due to a combination of the close proximity of the people in certain parts of the scenes (up to 25 cm) and their occlusion by other people. If people are positioned so that it is no longer possible to segment the foreground regions of different people in any view, POM is unable to detect all individual persons (as described in [6]).

Tracking Although the focus of the current paper is on person localization and track assignment, we also embed the results of both methods in a standard Kalman Filter (KF) framework, to compare results at the tracking level. We use a KF with a constant velocity model; the assignments of measurements are now made with respect to the KF predictions. We use a gating distance of 1.5 m to search for measurements from the locations corresponding to predictions. We require a track to be of certain duration, before it is considered active. Similarly, visible tracks are discontinued after a certain time during when no measurements are assigned. Both durations are set to 20 frames in the experiments. See Figure 4 (right, solid lines). As can be expected, the number of FP rises and the number of TP declines, when compared to the detection results (figure 4, right, dotted lines) which use GT data at time t. Nevertheless, the proposed method maintains its advantage versus the baseline POM method. Results can be seen in Figure 5.

Computational cost of both methods was assessed on a comparatively difficult 4-person sequences of 620 frames. Processing involved a single core Xeon 3 GHz system with 3 GB RAM. The POM detection method required about 7.5 s per frame, while our volume reconstruction took 3 s (both C++). The subsequent two-stage track assignment required 3 s per frame, for both localization approaches. This was reduced to 1.1 s when using 10 instead of 40 candidate assignments from preselection. All frames had a resolution of 752×560 pixels.

5 Conclusion

We presented an efficient two-step method for the joint person localization and track assignment in the context of a multi-view, multi-person tracking system. The proposed person localization approach, based on volume carving, outperformed a baseline POM localization method. This holds in particular for the

Fig. 5. Tracking sequences from four scenarios, one triplet of three time instances per scenario. Each triplet shows one of the three camera perspectives. Clockwise: (1) Four people starting a fight. (2) Three people argue. (3) People meet, hug and leave the scene. (4) people pass each other.

cases where people stand close together so that their projections are merged in the camera foregrounds. The POM method would converge onto the center of the cluster as the most likely person location; non-maxima suppression would discard the rest. Our system deals with this problem adequately.

References

1. Arsic, D., Hristov, E., Lehment, N., et al.: Applying multi layer homography for multi camera person tracking. In: ICDSC (2008)
2. Berclaz, J., Fleuret, F., Fua, P.: Principled detection-by-classification from multiple views. In: Proc. of CVTA (2008)
3. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
4. Calderara, S., Cucchiara, R., Prati, A.: Bayesian-competitive consistent labeling for people surveillance. *IEEE Trans. on PAMI* 30(2), 354–360 (2008)
5. Eshel, R., Moses, Y.: Homography based multiple camera detection and tracking of people in a dense crowd. In: *Proc. of the IEEE CVPR* . pp. 1–8 (2008)
6. Fleuret, F., Berclaz, J., Lengagne, R., et al.: Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. on PAMI* 30(2), 267–282 (2008)
7. Hu, W., Hu, M., Zhou, X., et al.: Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Trans. on PAMI* 28(4), 663–671 (2006)
8. Kang, J., Cohen, I., Medioni, G., et al.: Tracking people in crowded scenes across multiple cameras. In: *Proc. of the ACCV* (2004)
9. Khan, S., Shah, M.: Tracking multiple occluding people by localizing on multiple scene points. *IEEE Trans. on PAMI* 31(3), 505–519 (2009)
10. Leibe, B., Schindler, K., et al.: Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Trans. on PAMI* 30(10), 1683–1698 (2008)
11. Liem, M., Gavrila, D.M.: Multi-person tracking with overlapping cameras in complex, dynamic environments. In: *Proc. of the BMVC* (2009)
12. Mittal, A., Davis, L.: M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. *IJCV* 51(3), 189–293 (2003)
13. Murty, K.: An algorithm for ranking all the assignments in order of increasing cost. Operations Research 16(3), 682–687 (1968)
14. Szeliski, R.: Rapid octree construction from image sequences. CVGIP 58(1), 23 – 32 (1993)
15. Yang, D., Gonzalez-Banos, H., et al.: Counting people in crowds with a real-time network of simple image sensors. In: *Proc. of the IEEE ICCV*. pp. 122–129 (2003)