Multiple Camera Person Tracking in Multiple Layers Combining 2D and 3D Information

Dejan Arsić, Björn Schuller, Gerhard Rigoll

Institute for Human Machine Communication
Technische Universität München, Germany
arsic, schuller, rigoll @tum.de

Abstract. CCTV systems have been introduced in most public spaces in order to increase security. Video outputs are observed by human operators if possible but mostly used as a forensic tool. Therefore it seems desirable to automate video surveillance systems, in order to be able to detect potentially dangerous situations as soon as possible. Multi camera systems have seem to be the prerequisite for huge spaces where frequently occlusions appear. In this treatise we will present a system which robustly detects and tracks objects in a multi camera environment and performs a subsequent behavioral analysis based on luggage related events.

1 Introduction

Despite the legitimacy of a number of privacy issues, many systems have been deployed for surveillance applications. They generate large amounts of data that needs to be filtered out either for online detection of dangerous situations, or for offline information retrieval. Up to now these tasks are performed by human operators, of which a huge number is required if online analysis is required. Most of the data is stored in video archives without even being analyzed and is currently only used 'after the fact' as forensic tool, losing its primary benefit as an active real-time media. Therefore the demand for automated surveillance systems, providing a decision support interface to enhance the performance of a human operator, seems reasonable. This way unlawful acts could be detected in time or even prevented. This is not an easy task, as the system has to deal with large crowds resulting in severe occlusion, difficult and fast changing lighting situations, and views that are very narrow or too wide.
The observation of large spaces is mostly performed with a set of multiple cameras in order to handle occlusions of persons by other persons or objects. Besides technical issues such as time synchronization it is inevitable to correlate all camera views to analyze the scenery. Therefore a three staged approach has been implemented herein. First in all field of views (FOV) relevant regions are extracted with a foreground segmentation approach. The foreground is subsequently segmented into blobs, where the boundaries are determined. Next person tracking is performed by a homographic transformation [1] of the blob boundaries into the ground plane, where intersecting regions indicate person positions. Applying

the transformation in multiple layers provides a more precise 3D reconstruction of the scenery. Additionally incorrectly segmented foreground regions will not lead to drastic errors, as we perform a multi layer fusion. This method has a major drawbacks in scenes with overcrowding. In the first place false positives are frequently detected in scenes with multiple occlusions. Up to now there is no effective method to avoid these. Therefore we will will present an effective approach for false positive handling by transforming the resulting 3D surfaces into the corresponding 2D images and a further analysis. After the detection of so called object candidates a temporal analysis has to be performed. Applying a simple predictive Kalman filtering results in quite robust trajectories and consistent labels. Still occurring ID switches are prevented by additional tracking in the 2D data and the combination of the results.
In order to test and evaluate our algorithms we used the data presented in the 2007 workshop for Performance Evaluation of Tracking and Surveillance PETS2007 [2]. The data has been recorded near the check in area of an international airport from four different fields of view, which are illustrated in fig 1. During the recordings scenarios were played by actors while other passengers still could walk through the recording area, which created a realistic tracking environment with almost empty scenes and overcrowding. Four different scenarios, including loitering, leaving luggage, swapping luggage and stealing luggage have been recorded twice. With an additional scene with neutral behavior a total of 9 datasets is available.
This treatise is structured as follows: The basics of the homographic transformation and the extension to multiple layers will be discussed in section 2 and 3 followed by a novel approach for false positive handling in section. 4. The now known objects can be used for tracking, as described in section. 5 in combination with 2D tracking 6. We will conclude this work with an evaluation in section. 8 and a short conclusion in 9.

2 Object Detection

A major drawback in object detection with one single camera is occlusion handling. Two objects, which are occluding each other only partially will be recognized as one with common foreground segmentation methods [3]. Even methods such as KL Tracking [4] or color histogram based mean shift tracking [5] cannot solve the occlusion problem. Applying multiple camera perspectives provides the opportunity to solve the occlusion problem, as objects are probably not occluded from every perspective.
Therefore approaches using homographic transformation have been presented in [1] [6], which performs a transformation from one image plane into another one. A minimum of seven corresponding points in image and world coordinates have to be known to locate any blob position on the image plane, assuming, that the blob's lowest point is in contact with the ground. Therefore the transformation matrix has to be computed once for every single camera:

$$x' = Hx$$

Therefore the parameters h_{ij} of matrix H are computed with the Tsai camera calibration method [7], with 7 corresponding image points. This way we are able to transform object boundaries, found with a foreground segmentation algorithm based on Gaussion Mixture Models, into the ground plane of our world coordinate system or back into other fields of view. Fig. 1 illustrates the homography for all four fields of view. The transform can be basically interpreted as shadow on the ground plane created by a light source located at the camera position. The area within the outline of the polygon is considered as candidate for an object. As we cannot see through any object and determine the object's depth, the area can be quite long, especially if the camera position is as low as the one in the datasets third perspective.
Obviously there are some additional regions due to errors in the foreground segmentation task. These are not a major issue, as these usually do not appear in all camera views at the same time. Consequently these will be eliminated during the following fusion process.
Depth information is subsequently gathered by the computing intersections of

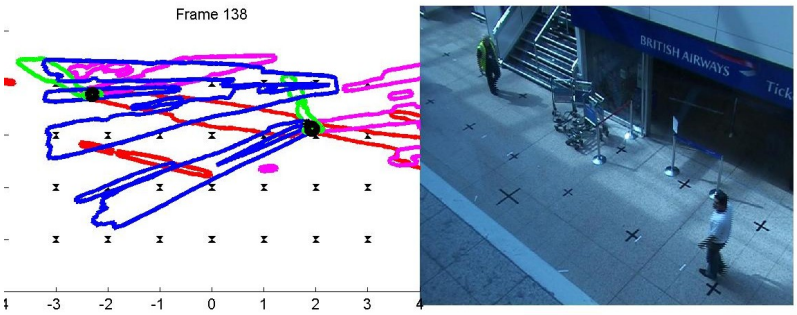

Fig. 1. Homographic transformation in Z=0 for all 4 views and subsequent fusion

all polygons in the plane, as shown in fig. 1. In all areas with more than three intersecting polygons object candidates can be assumed. Errors created by misclassification during the segmentation process should be eliminated, as it can be assumed, that segmentation is performed correctly in at least one camera perspective. Another strength of the homographic transform and the subsequent fusion is the robustness to occlusion. Each transformed blob denotes the region an object could be located in, even if it is not visible from the actual camera view. This way even objects not visible in any camera perspective could be detected. Actually it just denotes where a hidden object might be located. Therefore this areas are considered as object candidates.

3 Multi Layer Homography

Homography has been used just in the ground planein former works, as all pedestrians are usually located on the floor. Unfortunately results are quite unsatis-

fying because the feet are quite small objects and are frequently segmented incorrectly. In a real world scenario it has to be additionally considered that feet are not necessarily touching the ground while people are walking. As a result there might be a lack in precision of localization or some persons might not even be detected due to missing intersections. Even if there are intersections they would have to be grouped together as each feet would result in one separate intersection. An approach to group these small regions has been presented in [6] by assuming minimum sizes for objects, which would dismiss or mistakenly group regions. Therefore it seems reasonable to apply homography in additional layers. This way a pseudo 3D model of the object can be created and additionally the object's width and height can be estimated. Image 2 illustrates the reconstruction of a scene with an exemplary amount of 6 layers, where only intersections are shown. As it can be seen, both persons are perfectly separated and the person in front is recognized being taller than the person behind, which actually resembles the truth in this sequence. The 3rd person in the back is considered as background in this sequence as she is not moving for a while. The additional layers result in various sizes, which approximately resemble the sizes of the corresponding body parts. Especially the hips and the upper body lead to a better localization than just the feet. In a following step all layers are combined to one in the ground plane, which would result in a kind of a top view of the scene.

Computational effort can be kept at a minimum even for far more layers, as

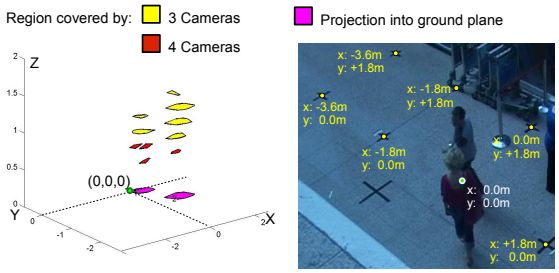

Fig. 2. Exemplary multi layer transform in 6 layers and projection into the ground plane

homography is only computed for 2 layers whereas all others can be computed by basic geometry, which can be computed simply by:

$$X_w(Z_w) = (X_1 - X_0)Z_w + X_0 \text{ and } Y_w(Z_w) = (Y_1 - Y_0)Z_w + X_0$$

Where (X_1/Y_1) and (X_0/Y_0) are the corresponding transformations in layers $Z_w = 1m$ and $Z_w = 0m$. The amount of layers highly depends on the difficulty of the scenario, but mostly 10 layers inbetween $Z_w = 0m$ and $Z_w = 2m$ seem sufficient, as most objects are shorter than two meters.

The multilayer approach produces very reliable results for small groups. For large groups and crowds the feet and heads areas usually do not show overlaps in the binary image, wheres the body area often touches. This frequently results in one

large connected object in the multilayer view. Assuming that there are mostly pedestrians in the scene, we can separate the resulting structure by analyzing the form.

4 False Positive Handling

The major drawback of the homography approach is quite frequent appearance of false positive candidate positions. This happens if two or more objects occlude similar background regions. The multilayer approach even leads to another 'floating' object. Unfortunately it is not possible to exactly determine whether there is an object hidden behind the other ones or a false positive has been created. Therefore we decided to detect and track so called false positive candidates.
These are detected by transforming the detected regions back into the 2D images. As the 3D information is now available, we know which object candidate is located occluded by another one and compare their sizes. In a subsequent step a heuristic approach is applied: if an object candidate is visible at least from one camera perspective, that means it is the first one in a row or larger than the objects occluding it, it is considered as a real one. Being totally covered in all camera perspectives indicates a false positive candidate. These are still used for object tracking, as it might also be a hidden object. Therefore new appearing objects are associated with the false positive candidate if tracking criteria (see sec. 5) match, or an object disappears and the false positive candidate appears at more or less the same spot.

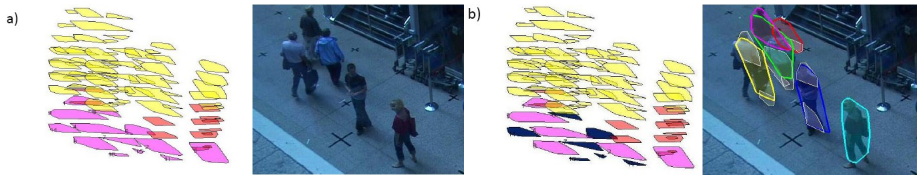

Fig. 3. a) The reconstructed scene with false positives. b) After false positive detection

An exemplary result is shown in fig.3 where the projections into the ground plane have been transformed back into 2D. In fig. 3a) the pink regions represent all candidate positions found in the image. Obviously there are more candidates than persons in the image. Image 3b) illustrates the back projection of all areas where the white areas in the image are false positives and the others represent detected objects. The false positives can then be removed from the reconstructed image, here in black.

5 Object Tracking

Up to now the system is only able to determine the *xy-position* in world coordinates. Unfortunately there is no temporal information about the detected

regions. Therefore a system creating relationships in between regions in an image sequence is required. This has been implemented applying a two staged approach, where a Kalman Filter [8] is used as basis in order to predict an object's position x_k in the next frame x_{k+1}using a motion model:

$$x_{k+1} = F_{k+1,k} * x_k + w_k$$

In a subsequent step a measurement is performed to find the object fitting best to the prediction:

$$y_k = H_k + x_k + v_k$$

One of the main error sources are elements appearing or disappearing just for

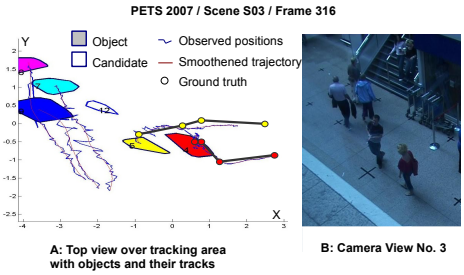

Fig. 4. a) Exemplary trajectories in world coordinates. b) Frame from sequence

a few frames. For instance a person could disappear for a few frames and a new object is appearing in the neighborhood. Or it may simply be dismissed and after reappearance be assigned a new ID. Although the predicted position and the measurement do not fit exactly, the new object might get the ID of the vanished object, even if it reappears after a few frames. It seems reasonable to take the distance of the object positions into account and memorize old object positions. Therefore a prediction-observation likelihood

$$L = logP(y_k|y_{1:k-1})$$

is computed evaluating a Gaussian density [6]. If a reappearing object has a smaller distance, a higher probability and additionally matches a prediction of the Kalman filter the the older object ID gets the original ID reassigned and the new object gets a newer one. The computed trajectories are illustrated in fig. 4

6 Object Recognition with SIFT Features

While Multilayer Homography and Adaptive Foreground Segmentation are efficient tools for tracking even in moderately crowded areas, there are a number of situations (e.g. occlusions) where we may inadvertently confuse two tracked persons. A typical example would be two persons disappearing behind the same

obstacle. Once they emerge from the said obstacle, we need a way to ascertain their identity before continuing the tracking process. Another scenario would be one person entering the scene through a door as another tracked person disappears from the scene through the same door. Again we would need an indicator that the person entering the scene is not identical to the person leaving the scene, even if they appeared in the same spot.
For our identification procedure we use features obtained by scale invariant feature transform (SIFT). SIFT was introduced in 1999 by Lowe and uses scale- and rotation-invariant descriptors derived from gradient-histograms at scale-space extrema to find unique features in a scene [9].
These features, especially when recorded over a sequence of frames, can be used to resolve tracking ambiguities. The approach is similar to SIFT face-recognition as described in [10] and [11]. However, instead of faces we look at the whole body of the person to be tracked. We also have to consider swinging motions of arms and legs, as well as partial occlusion by obstacles or body movements from frame to frame.
To account for these difficulties, we record SIFT features and their relative positions over several frames, effectively tracking the person by their SIFT features. The recorded data can be thought of as a graph, where the nodes are specific SIFT descriptors and the edges describe their relative spatial position. When an ambiguity occurs, we check the SIFT features of the possibly switched persons against the recorded SIFT graphs by non-rigid matching. A person's graph M_{Person} can be described as follows:

$$M_{Person} = \{D^M, P^M\}$$

where D^M signifies the vertices, i.e. the 128-element descriptor vector, and P^M signifies the edges, i.e. spatial positions of the feature.

To reliably build a SIFT model of a person, we track the person's SIFT features over as many frames as possible while recording the found features and their positions in each frame. The person's temporary model consists of two groups of SIFT features: those which were repeatedly detected over several frames in similar locations (tracking group, Ω_T) and those which where so far detected only once or twice (candidate group, Ω_C). Since we have to account for the changing appearance of the moving person, we use a scoring system for updating both groups. Features which are not detected for some time are gradually forgotten, while repeated detection strengthens a feature. This is achieved by assigning numerical weights to each feature, which are adjusted in each frame. Features in Ω_C which are detected often enough are eventually promoted to Ω_T. This approach is similar to [12], but uses a simplified mechanism. This leads to

$$\Omega_I = \{D^{\Omega_I}, P^{\Omega_I}, W^{\Omega_O}\}$$

with D^{Ω_I} as descriptor vectors of features, P^{Ω_I} the spatial positions within reference frame and W^{Ω_I} is the forgetting factors for descriptors for $I \in C, T$
A candidate Ω_C is added to the tracking group Ω_T if W^{Ω_C} is larger than a predefined promotion threshold α_C. The corresponding SIFT features $D_i^{\Omega_C}$ are

then added to Ω_T and removed from Ω_C. Otherwise, if $W_j^{\Omega_T} < \alpha_T$, the SIFT feature $D_j^{\Omega_T}$ is added to Ω_C and removed from Ω_T.
In each frame, we search for features of the tracking group around their expected positions. The set of found features is refined by an iterative elimination of bad matches based on relative distances (with compensation for a changing scale). In a next step, known and new features of the candidate group are found and strengthened. All found features of the tracking group are recorded continuously in a separate group for the construction of the average SIFT model. Once the tracking on the person is lost, the recorded features are searched for stable, reoccurring features: we group similar SIFT descriptors (i.e. Euclidean distance smaller than defined threshold) with small spatial distances and use their average position and descriptors to create an individual graph M_{Person} (nodes are SIFT descriptors, edges are spatial positions) modeling the person tracked. Weak features, i.e. features which were detected in a region only a few times, are discarded. This reduces unreliable fast-moving elements like arms and legs to a considerable degree.

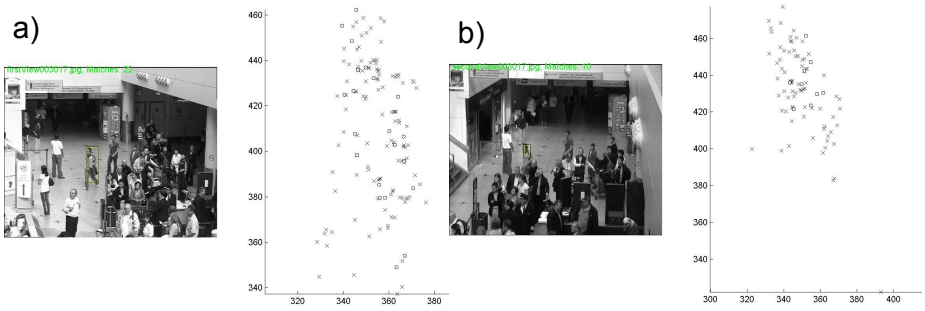

Fig. 5. Tracked person with bounding box in 2 views with matches P^{Ω_T}

To solve ambiguities and avoid miss-identification, the persons involved are matched against the respective SIFT models. The person is non-rigid, may have turned slightly and features may be obscured or not visible at all, requiring a flexible matching algorithm. Matching from $M_{Person(i)}$ to $\Omega_{C)}$ is achieved by minimizing the energy function

$$E(x)_{ij} = x_{ij}^T H_{ij} x_{ij} + c_{ij}^T x_{ij}$$

with x_{ij} the binary vector containing possible feature assignments, H_{ij} a matrix containing respective spatial distances and c_{ij} the vector of Euclidean distances between descriptors, by means of quadratic binary programming (since it is a quadratic assignment problem). To avoid calculating every possible combination of features, we filter by the Euclidean distances of the descriptors. As the problem is NP-hard, we use an approximation based on [13]. We select the assignment x_{ij} with the lowest matching energy E_{ij} and highest number of positive feature matches as the match for an individual person. Therefore we can

assign the track of model M_P to the person with Ω_C and can thus resume the tracking. Since in a crowded scene important SIFT features may be obscured and non-rigid matching is also non-robust, we repeat this process in several frames to achieve a reliable match. The tracking and identification described above works well as long as the person does not turn too fast. However, since several points of view are available, we can observe the same person from different angles. A person visible to four cameras is therefore described by four SIFT graphs attributed to that person (see (5)). By comparing the SIFT graphs from the different views, we can detect turning motions by the occlusion and re-appearance of features and re-identify the person by the known SIFT models as outlined above.

7 Event Detection

Due to the lack of a complex database we decided to work with an expert knowledge based approach. Instead of creating a probabilistic model, a set of rules has been defined to detect events. According to the definition provided for the PETS 2007 challenge a person is loitering if he stays in the field of view for at least 60 sec. This task can be solved rather easily, by adding a time stamp $LifeTime$ to every tracked object if it appears for the first time. If the difference is larger than the required lifetime within the camera view, an alert can be created.
Having a close look at the provided data sets, it becomes clear, that leaving luggage is following a quite similar scheme. In the first place an object will split into two. One of these parts will remain stationary in the following frames. The other one will be leaving the stationary one after a while. A warning is displayed if the object distance is larger than $3m$. If the distance is larger than $3m$ for at least $30s$ an alert signal is displayed. Figure 6 illustrates this process, which has been implemented as a simple decision tree. A second case is that a split is detected and after a short while the moving object vanishes within the $3m$ region. This commonly happens near the borders of the field of view. Therefore the sudden vanishing of the moving split object has to be also modeled.

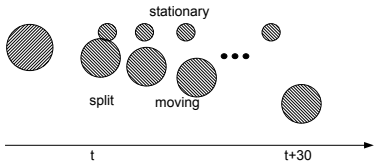

Fig. 6. Scheme of leaving a piece of luggage

8 Evaluation

For each surveillance system it is crucial to provide some measurements for evaluation. Yin et al [14] have proposed a variety of measurements of tracking systems based on their trajectories. For first evaluations we decided to determine the Euclidean distance between the labeled position and the detected one and count

ID changes when only homography is used. The evaluation will be explained exemplary for Scene $S03$. Figure 7 shows exemplary the Euclidean distance for both labeled persons in the sequence, based on the detected IDs. In the first row results for $ID1$ are given. The different colors of the graph indicate ID changes. Here at frame 550 with $ID2$ and at frame 1600 with $ID3$. Leaving swaps aside a high precision of localization is achieved using only homography. In contrast $ID2$ changes its ID only once during tracking and shows also a high accuracy of about $12cm$. Compared with the results from [**?**] the error could be reduced by $13cm$

As labeling has been performed manually there is obviously some discrepancy

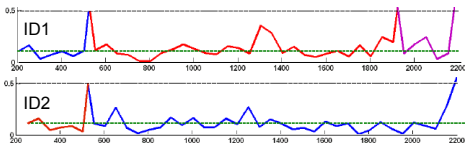

Fig. 7. Euclidean distance for IDs 1 and 2 in S03 for homography only. Color changes indicate ID swaps.

between ground truth and real position. During tracking the estimated center of the projection is used as xy-position in world coordinates, whereas ground truth is often provided for outer boundaries of the persons. Additionally, calibration errors have to be taken into account. Therefore an average distance of $13cm$ with a small variance should be fully satisfying. Image 8 shows an example for positioning of the bounding boxes in the corresponding image data.

Additional identification with SIFT Features lowers the number of ID changes

Fig. 8. Corresponding bounding Boxes in the camera views

Scene:	t_d	t_e	δt	x	y
S01	1648	148	23	-0.730388	0.845811
S02	1718	218	89	-0.161693	-0.172595

Table 1. Timestamps and positions for loitering

from 15 to only 2 in the PETS2007 dataset without affecting the accuracy of positioning.Especially the false positive handling enhances the tracking performance drastically as by far lower candidates are detected. The number of false

positives could be reduced by 72% to a value of 24 in ll 8 sequences of the PETS2007 data set. These only occured for a short time period and did not affect tracking.
The event detection task profits from the enhanced tracking part as less false alerts are occurring. Compared to previous results, detected timestamps and positions almost do not differ. Table 1 shows the frame number for detection of loitering t_d and a fairly small time difference δt ground truth measure. There were no misses and only one insertion, as a bag was stationary in the video for a longer time, which could be interpreted as clue for an unruly event.
The recognition of luggage events is by far more complicated than the detection of loitering people. Especially the split and merge detection tends to be rather difficult. The maximum distance between two objects has to be set carefully and be fitted to each application scenario. Here the maximum distance for splits was set to $0.5r$ and resulted in no false positive and no miss for left luggage detection. Table 2 shows both the time stamps for the detection of unattended luggage and abandoned luggage. Additionally the position of the original owner (x_{own}, y_{own}) is indicated, if available.The presented system has been additionally tested on the PETS2006 set with similar results, although the camera setup was quite different.

Scene:	t_d	x	y	x_{own}	y_{own}
S07 unatt.	1491	-0.01	-0.20	NA	NA
S08 unatt.	1147	-0.14	0.04	-2.38	-1.05
S08 left	1773	-0.12	0.04	NA	NA

Table 2. Timestamps and positions for unattended and left luggage in S07 and S08. The position of the owner is also given, if available

9 Conclusion and Outlook

In this work we have shown extensions to the well known tracking approach using homographic transformation. Tracking performance has been drastically increased by simple methods. In the first place additional precision has been added to localization performance by adding detection steps in multiple layers, which led to a almost 50% smaller error of $13cm$. This especially helped when feet were incorrectly segmented or not touching the ground plane. Additionally the false positive rate has been drastically reduced by a further analysis of the resulting regions. These are subsequently transformed into the 2D images.
In future works there has to be a discrimination between objects and human beings to improve the event detection performance. Pedestrian detection systems have already been presented in [15]. Unfortunately these systems rely on training data, which is only available for frontal views, and even function with some robustness against occlusion. An object's texture could be used to create a pseudo 3D model. This might increase tracking performance drastically after

splits and merges of objects. In following works the popular Kalman filter could be replaced by a predictor with a nonlinear motion model such as the unscented Kalman filter [16], as human motions can change drastically.
For a more detailed behavior detection it would be reasonable to further choose the best sight of the person and perform a analysis based on motion features.

References

1. Khan, S.M., Shah, M.: A multiview approach to tracking people in crowded scenes using planar homography constraint. In: In Proceedings IEEE Conference ECCV 2006. (2006) 133–146
2. Ferryman, J., Tweed, D.: An overview of the pets 2007 dataset. In: Proceeding Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS 2007, IEEE, Rio de Janeiro, Brazil. (2007)
3. Stauffer, C.: Adaptive background mixture models for real-time tracking. In: Proc. of IEEE conference on Computer Vision and Pattern Recoginitions. (1999) 246–252
4. Shi, J., Tomasi, C.: Good features to track. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94), Seattle (1994)
5. Zivkovic, Z., Krose, B.: An em-like algorithm for color-histogram-based object tracking. In: Proceedings IEEE Conference CVPR 2004. (2004)
6. Arsić, D., Hofmann, M., Schuller, B., Rigoll, G.: Multi-camera person tracking and left luggage detection applying homographic transformation. In: Proceeding Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS 2007, IEEE, Rio de Janeiro, Brazil. (2007)
7. Tsai, R.Y.: An efficient and accurate camera calibration technique for 3d machine vision. In: Proc. of IEEE conference on Computer Vision and Pattern Recoginitions. (1986) 364–374
8. Kalman, R.E.: A new approach to linear filtering and prediction problems. Transactions of the ASME Journal of Basic Engineering (1960) 35–45
9. Lowe, D.G.: Object recognition from local scale-invariant features. iccv **02** (1999) 1150
10. Kisku, D., Rattani, A., Grosso, E., Tistarelli, M.: Face identification by sift-based complete graph topology. Automatic Identification Advanced Technologies, 2007 IEEE Workshop on (2007) 63–68
11. Luo, J., Ma, Y., Takikawa, E., Lao, S., Kawade, M., Lu, B.L.: Person-specific sift features for face recognition. Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on **2** (2007) II–593–II–596
12. Tang, F., Tao, H.: Object tracking with dynamic feature graph. Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on (2005) 25–32
13. Berg, A., Berg, T., Malik, J.: Shape matching and object recognition using low distortion correspondences. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on **1** (2005) 26–33 vol. 1
14. Yin, F., Makris, D., Velastin, S.: Performance evaluation of tracking algorithms. In: Proceeding Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS 2007, IEEE, Rio de Janeiro, Brazil. (2007)
15. Papageorgiou, C., Poggio, T.: A trainable system for object detection. International Journal of Computer Vision **38** (2000) 15–33
16. Wan, E., van der Merwe, R.: 7. In: Kalman Filtering and Neural Networks. Wiley (2001)