Application-driven merging and analysis of person trajectories for distributed smart camera networks

Jürgen Metzler[a], Eduardo Monari[a], and Colin Kuntzsch[b]

[a]Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Fraunhoferstrasse 1, 76131 Karlsruhe, Germany;
[b]Kartographie und Geoinformatik, Leibniz Universität Hannover, Appelstrasse 9a, 30167 Hannover, Germany

ABSTRACT

Tracking of persons and analysis of their trajectories are important tasks of surveillance systems as they support the monitoring personnel. However, this trend is accompanied by an increasing demand on smarter camera networks carrying out surveillance tasks autonomously. Thus, there is a higher system complexity so that requirements on the video analysis algorithms are increasing as well. In this paper, we present a system concept and application for anonymously gathering, processing and analysis of trajectories in distributed smart camera networks. It allows a multitude of analysis techniques such as inspecting individual properties of the observed movement in real-time. Additionally, the anonymous movement data allows long-term storage and big data analyses for statistical purposes. The system described in this paper has been implemented as prototype system and deployed for proof of concept under real conditions at the entrance hall of the Leibniz University Hannover. It shows an overall stable performance, particularly with respect to significant illumination changes over hours, as well as regarding the reduction of false positives by post processing and trajectory merging performed on top of a panorama based person detection module.

Keywords: video surveillance, distributed smart camera networks, trajectory analysis

1. INTRODUCTION

Several approaches for gathering and analysis of person movements in distributed multi-camera networks have been developed in recent years.[1–5] However, this trend is accompanied by an increasing demand on smarter camera networks carrying out surveillance tasks autonomously. For instance, it is desirable that cameras slew autonomously in order to get an overlap of the camera fields of view so that an individual can be hand-off from camera to camera. In general, using non-static cameras allow a better coverage of surveillance areas without the need of additional hardware. Unfortunately, at the same time the requirements on algorithms are increasing as well, due to the higher complexity of video analysis.

In this paper, we present a system concept and application for gathering, processing and analysis of trajectories in distributed smart camera networks. The system is split into three level of data processing:

On a first level (signal level), an approach for person detection based on background subtraction is applied. Although classical detectors assume a static camera, the presented approach has been enhanced to be used for non-static cameras as well. In order to do this, the images from a video stream are registered on each other first. The result is a joint static panorama image that represents the learned background. Once the background panorama is generated, foreground (motion) segmentation is achieved by a real-time frame-to-panorama registration, followed by a classical background subtraction. After motion segmentation, connected foreground regions are detected and analyzed. They are verified for persons by analyzing their geometric properties.

On a second processing level (feature/object level), a single-camera tracking approach is presented. The tracking is performed locally by individual cameras embedded within a dynamically reconfiguring camera network. It performs a multi-object-tracking together with an explicit occlusion handling. A continuous tracking of persons across merge & split situations is important for trajectories analysis. It is solved through a merge & split detection process prior to the update of the tracks. Whenever two tracks have been registered as merged, the tracking module keeps track of the merged tracks maintaining the number of the persons inside the joint track.

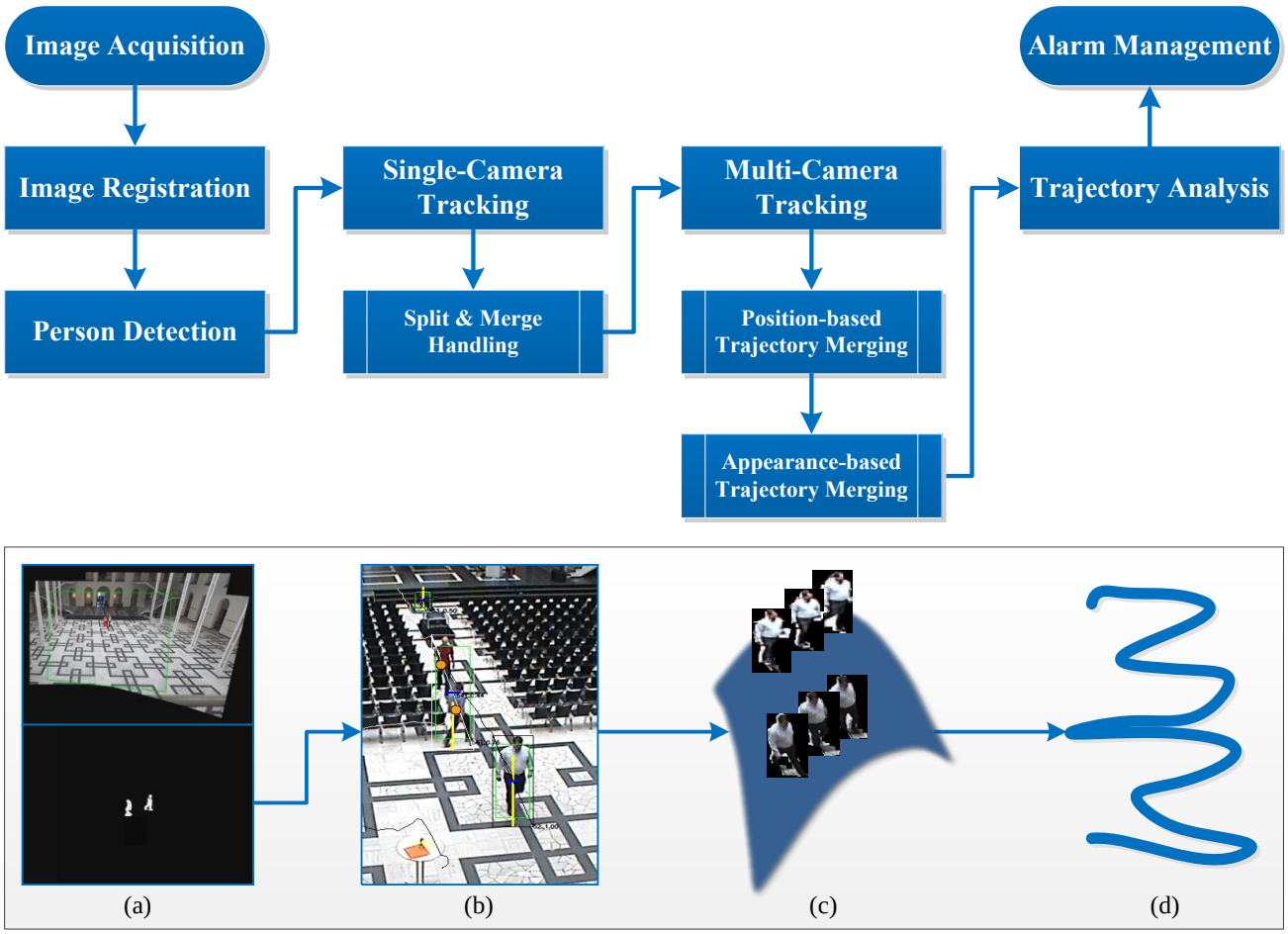


Figure 1: Flow chart of the proposed system (sub-processes are marked by white vertical lines): (a) image registration and person detection - motion-based detection of persons in video streams of non-static cameras, (b) single-camera tracking - gathering of reliable partial trajectories, so-called tracklets, (c) multi-camera tracking - position- and appearance-based merging of tracklets and (d) trajectory analysis - detection of predefined or rare trajectories.

If a split is detected, the persons are reassigned to the corresponding single tracks again. Thus, anonymous identities can be retained so that there are no gaps during merge situations which is an important advantageous for the trajectory analysis.

Finally, at the third level (mulit-camera / object level), the post-processing of single-camera trajectories to global ones by position- and appearance-based fusion of them is proposed. Individual tracks from single-camera tracking, that are naturally spatially restricted to a single camera field of view, need to be integrated into common trajectory knowledge about tracked objects moving within the camera network. Adding multi-camera tracking capabilities to the system allows active tracking and re-identification of persons through the camera network by collaborative reconfiguration of pairs of cameras in order create a single trajectory spanning a spatially extended area covered by multiple fields of view. A purely geometric approach exploits properties of the spatio-temporal relations between trajectory data observed by different sensors within the camera network. By employing an appearance-based model for observed persons on top of our first solution, we are then capable of reliably tracking persons through the camera network, resulting in reliable global trajectories.

2. SYSTEM OVERVIEW

The main objective of the system is to gather trajectories of persons and to detect conspicuous ones in near real-time. At that, trajectories which shall be detected, can be ones that are previously defined or just rarely appear. In the following, the main modules of the system shown in Fig. 1 are described.

Person Detection

We use a motion-based detector as we only interested in the analysis of the movements. Although the detector assumes a *static* background, it is suitable for non-static cameras. In order to do this, the images from a video stream are registered on each other first. The result is a panorama image that represents the *learned* background and that then can be used for foreground (motion) segmentation. After motion segmentation, connected foreground regions are detected and the dimensions of these so-called blobs are analyzed: they are verified for persons using calibration information.

Single-Camera Tracking

The tracking module performs a multi-object-tracking updating the tracks with the new blob information in every frame of one video stream. As we are interested to detect unusual trajectories that not be necessarily stick to a certain dynamic model, the tracking does not implement a state estimation with a dynamic model but performs a nearest-neighbor assignment of blobs to tracks in each frame. When two or more persons come close to each other, occlusion effects arise. In this case, the detection step only extracts one blob for the persons. The tracking module therefore has to explicitly deal with merge & split effects of the underlying blobs. This is solved through a merge & split detection process prior to the update of the tracks. Whenever two blobs have been registered as merged, the tracking module keeps track of the merged blob maintaining the number of the persons inside the joint track. If a split is detected, the persons are reassigned to the corresponding single tracks again. Thus, a continuous tracking of persons across merge & split situations can be achieved. The split handling can thereby subdivide the previous track into a number of new tracks considering the new blobs and the last known positions of the persons.

Multi-Camera Tracking

As the tracking task is performed locally by individual cameras embedded within a dynamically reconfiguring camera network, there can be a gap between the camera fields of view or these areas may spatially overlap occasionally. Individual tracklets from single-camera tracking, that are naturally spatially restricted to a single camera field of view, need to be integrated into a common trajectory knowledge about tracked objects moving within the camera network. Adding multi-camera tracking capabilities to the system allows active tracking and re-identification of persons through the camera network by collaborative reconfiguration of pairs of cameras in order create a single trajectory spanning a spatially extended area covered by multiple fields of view.

Trajectory Analysis

Our design of retaining tracking capabilities without biometric techniques allows automatized, real-time collection of large (and wrt. to the natural limitations of coverage, complete) anonymous trajectory data. This type of person movement data can easily be used for various trajectory analysis techniques of varying complexity in contexts like statistical data collection, as data is available and storable for extended periods of time, or security related applications, in which real-time analyses (e.g. anomaly-detection) make security-relevant information available to e.g. security personnel.

3. PERSON TRACKING

In our proposed approach, we use a single-camera multi-object tracker based on a foreground-background segmentation (motion segmentation). It updates tracks with new blobs (connected foreground regions) in every frame of one video stream. In order to do this while the camera is rotating, an image-to-panorama registration is used for motion segmentation.

3.1 Person Detection

Person or motion detection is basically performed by background subtraction and blob validation as post processing. One advantage of background subtraction is the simultaneous object segmentation, which is of high interest e.g. for appearance feature extraction of objects. However, for non-static cameras (as used in this work) background subtraction is a quite challenging task, since over time, pixel correspondences can become invalid. The approach used in this work is based on the idea to generate a panorama background model (by image stitching/mosaicing) and to perform background subtraction and update only for the part of the panorama image where the current image frame is located. In the next subsection the image-to-panorama registration is presented before then an adapted background estimation and subtraction strategy of the motion segmentation for pan-tilt cameras is presented.

3.1.1 Image-to-Panorama Registration

The basic component of the whole approach is a real-time image-to-image registration with sub-pixel accuracy called m^{3}motion.[6] This algorithm allows for an efficient and reliable estimation of the homography of two sequential video frames.

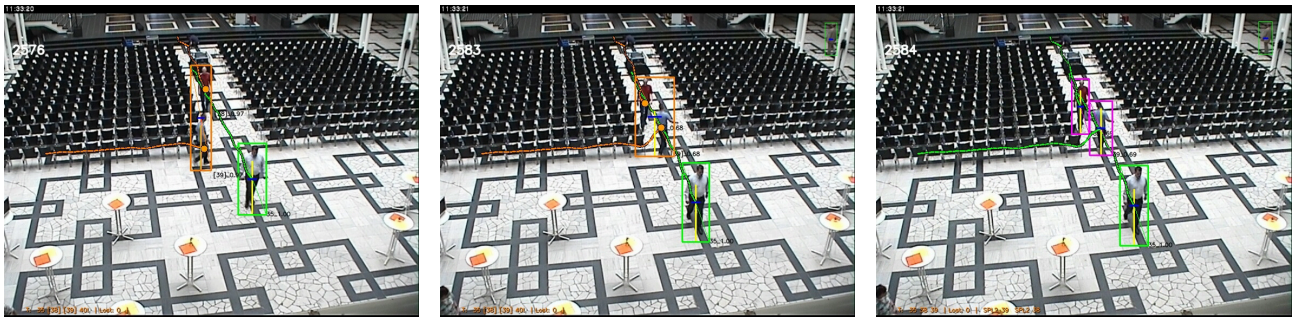
Given the homography of two sequential images $\mathbf{H}_{t_k,t_{k-1}}$ for each normalized pixel coordinate $\mathbf{p} = (x, y, 1)^T$ in the current video frame $I(t_k)$, the calculation of the corresponding coordinate $\mathbf{p}' = (wx', wy', w)^T$ in $I(t_{k-1})$ can be determined by $\mathbf{p}' = \mathbf{H}\mathbf{p}$. In this case, the video frame $I(t_k)$ is transformed into the coordinate system of frame $I(t_{k-1})$. However, for generation of a large panorama image, the goal is to warp all video frames into a common pixel coordinate system. But unfortunately, in most cases it is not possible to define a single reference image with a sufficient overlap to all sequential video frames for homography estimation. To achieve a image-panorama registration in[7] a two step process has been proposed. The process starts with an initialization step for generation of a so-called key-frame map. The key-frame map is a representation of a panorama image by a small number of video frames $\{I_0^{key}, I_1^{key}, I_2^{key}, \ldots, I_G^{key}\}$ with corresponding reference-homography matrices $\{\mathbf{I}, \mathbf{H}_1^{key}, \mathbf{H}_2^{key}, \ldots, \mathbf{H}_G^{key}\}$ to a predefined initial key frame $\mathbf{H}_0^{key}$. During this first phase, no background estimation and subtraction is performed. After generation of the key-frame map, background estimation and subtraction is performed. The background panorama is generated by registration of each following video frame into the reference pixel map. This can be achieved, by a frame-to-frame homography estimation between current video frame and the key-frame with the largest overlap to the current frame. As a quality coefficient for the frame-to-frame correspondence, the resulting image overlap and total number of found corresponding feature points is used.

The homography between the current video frame and the reference pixel map can now be calculated by multiplication of the homography $\mathbf{H}_{t_k,s}$ between current video frame $I(t_k)$ and key-frame I_s^{key} , with the reference-homography of the key-frame $\mathbf{H}_s^{key}$: $\mathbf{H}_{t_k,t_0} = \mathbf{H}_{t_k,s}\mathbf{H}_s^{key}$. This step is equivalent to a "frame-to-panorama"-registration which allows for generation of a background panorama in a reference pixel coordinate system, as well as motion detection using an arbitrary background subtraction approach.

3.1.2 Background Estimation and Subtraction

For pixel-based background estimation an enhanced version of the $\Sigma\Delta$-approach introduced by A. Manzanera et al. in[8] is used. It models each pixel as a single Gaussian intensity distribution and is characterized by very low memory and computational costs. For the purpose of panorama based background estimation and subtraction, we enhanced the $\Sigma\Delta$-approach by a binary mask. It allows to distinguish between pixels of the background model that are visible in the current frame of the pan-tilt camera, and pixels that are not visible temporarily.

From a probabilistic point of view, we can then assume that for pixels that are not temporarily visible, the uncertainty of intensity variance is increasing. To limit this, we additionally introduce a maximum pixel variance to avoid unreasonable high threshold level in cases where certain pixels of the panorama are not visible for a longer period of time. However, the maximum variance is chosen in a way that the resulting threshold is high enough to avoid false positive detections immediately after slewing the camera to an out-dated region of the panorama background model.



(a) (b) (c)

Figure 2: Example of a resolved merge & split situation and results of the localization refinement (orange circles): (a)-(b) detected merge situations (big orange rectangle) and (c) detected split (darker/cyan rectangle).

For our purpose of panorama-based background estimation and subtraction, the modeling of the uncertainty leads to suitable handling of a self-adapting thresholding and to an adaptive background update speed depending on the uncertainty. For more details on background estimation and update strategy, as well as dynamic thresholding of this approach, please refer to.[8]

3.2 Single-Camera Tracking

The single-camera person tracking is performed by a multi-object-tracking updating the tracks with the new blobs from the person detection module in every frame of one video stream. New blobs are associated to the tracks using their positions in the images which is on the one hand an easy and a reliable process if there are only few persons in the scene. On the other hand, an increasing number of persons makes this task more difficult as the cases in that several persons come close to each increase as well. In such cases, the person detection step only extracts one blob for the persons. This is solved through a merge & split handling that automatically detects merges and splits based on the number, position and dimension of the blobs. Also, a localization refinement that use estimated person heights and densities in the images to determine the positions of the persons in a merged blob is applied.[9] Thus, trajectories across merge & split situations can be achieved (see Fig. 2 for an example). Also, as for the trajectory analysis it is important to get reliable trajectories, we compute a confidence measure that is considered in the trajectory analysis.

3.3 Multi-Camera Tracking

The aim of multi-camera tracking is to gather as complete person movement information as possible within the smart camera network, taking into account incomplete coverage of the observed area and, consequently, continuous camera movement. In the context of our work, we deliberately avoid biometrics-based re-identification techniques. In this section, we present two approaches for constructing long trajectories spanning multiple camera fields of view while retaining the anonymous identity of observed persons. The first approach exploits properties of the spatio-temporal relations between trajectory data observed by different smart cameras within the network. After discussing qualitatively the assumptions we make about the data provided by multiple instances of the single-camera tracker, taking into account uncertainty and errors, we present a second approach to multi-camera tracking: we augment the purely spatio-temporal trajectory matching process by employing an appearance-based model for observed persons.

3.3.1 Position-based Merging of Trajectories

For the task of integrating the continuous stream of tracklets from the distributed camera network, we employ a centralized data aggregation module. It identifies sets of two or more tracklets from different sensors that correspond to observations of the same person's movements through the camera network. Output of this module is a real-time stream of trajectories which relates 1:1 to persons currently observed (effectively removing the redundancy of the tracklets while mapping tracklets to person's complete trajectories).

The process is based on the following assumptions about the sensor network and the single-camera tracker. Most importantly, we assume time-synchronization between the cameras within the distributed camera network in order to be capable of match the distributed observations by timestamp. In this regards, we also assume that the results of the single-camera tracker are available for data integration almost simultaneously. This assumptions can be bypassed with data buffering leading to a systematic delay between input and output. Hereby, the delay would be in the order of its magnitude resulting from the different times of availability of single-camera tracker results. Also, the algorithm does not attempt to correct the results of the single-camera tracker. It is, however, auto-correcting in its ability to detect false matches between tracklets on-the-fly (see Fig. 3e). First step of the process is to look for spatially corresponding observations made at the same time (thus the synchronicity requirements, see Fig. 3a and Fig. 3b for spatial correspondency with and without temporal correspondency). The idea is to eliminate all pairs of potential matches for which observations at the same time were made in significantly different places (in this case we conclude that the respective observations correspond to different persons), using a parameter representing the worst-case spatial resolution of the single-camera tracker. All observations are only compared to observations from different cameras (under the assumption that separate results provided by the single-camera tracker imply separate entities). Result of this step is a number of sets of spatio-temporally corresponding observations, each containing at least a single observation from a single sensor or a number of observations from different sensors with overlapping fields of view. For each of these corresponding sets we report a representative position as a the first point of a global trajectory.

For observations in subsequent iterations, i.e. for each update with more recent sets of observations from the camera network, one of three possible cases occurs:

1. The observation is part of a tracklet that is part of an already existing correspondency group (i.e. has already been part of a global trajectory previously). In this case, it simply remains to be checked whether the correspondency with other tracklets from different cameras, which has been identified at an earlier point in time, is confirmed by more recent observations from all of the involved tracklets. If this is the case, we simply continue using the set of corresponding observations to generate a representative location for the global trajectory at the given time (this includes cases where an observation has no correspondencies from other cameras, i.e. the person is currently observed by only one camera, see Fig. 3a). Otherwise, i.e. as soon as we notice the spatial correspondency no longer holds, we need to split the previously merged trajectory into subsets of corresponding observations, effectively creating a second global trajectory as soon as the process notices the false match between observations from an earlier iteration (see Fig. 3e).

2. The observation has not yet been integrated into a global trajectory previously, i.e. it is the start of a new tracklet. In this case, we continue similar to the initial step of finding spatio-temporal correspondencies with other simultaneous observations. If the identified correspondencies are already part of a global trajectory, we add the new observation to its list of correspondencies. Otherwise, we start a new global trajectory. Observations that have been assigned to different correspondency groups, i.e. pairs of observations that were rejected as candidates for matching, are under no circumstances matched at a later point in time (see Fig. 3d).

3. A global trajectory created at a previous iteration is no longer updated by more recent observations. In this case, we can remove it from the list of global trajectories after a certain timeout. Within this timeout missing information does not cause conflicts with previous matching results (see Fig. 3c).

3.3.2 Appearance-based Merging of Trajectories

The objective of the appearance-based merging of trajectories is to merge tracklets of the same individual gathered from different cameras at the same time. Another one is the linking of tracklets of the same individual that appeared at different locations in the same camera field of view (e.g. after occlusions). However, there are several challenges which makes this task difficult such as significant changes of appearance of persons, different illumination or camera parameters. In addition, for instance, in surveillance scenarios only low-resolution videos are usually available, so that biometric approaches may not be applied.

In our system, we apply a whole-body person re-identification approach for appearance-based merging and linking of tracklets that is suitable for low-resolution videos.[10] The method is divided in two stages: first, an appearance model is computed from several images of an individual and pairwise compared to each other. The

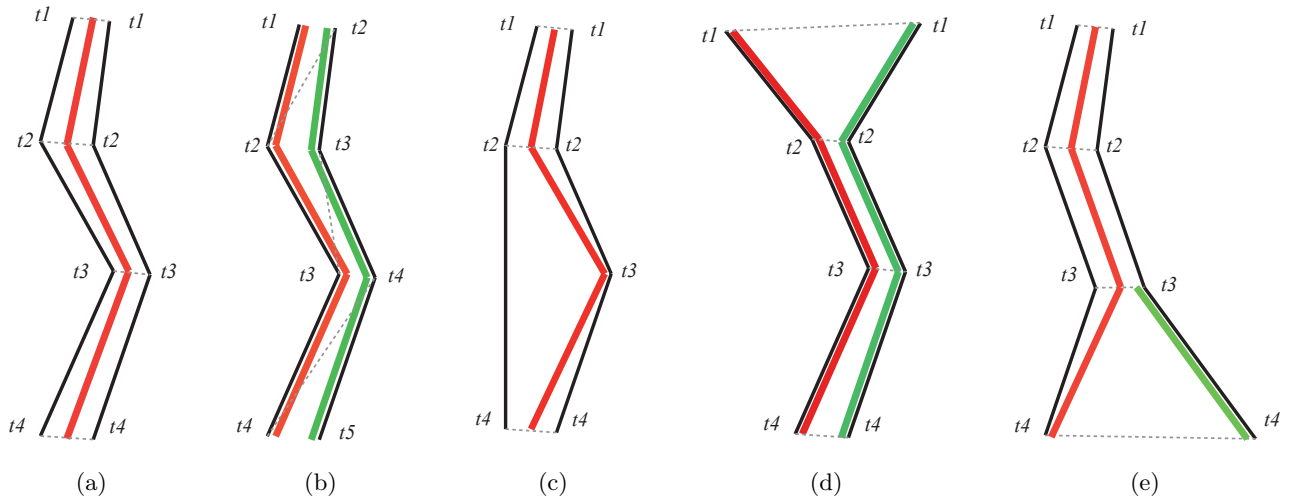


Figure 3: Typical situations occurring within the multi-camera tracker. Input is provided by individual cameras as stream of single observations (points with timestamps, exemplary called t1-t5 in chronological order). For each observation, all spatio-temporally corresponding observations from other cameras are identified by spatial matching of temporally corresponding observations by their synchronized timestamps (depicted as dotted line between temporally corresponding observations). On top of the original input, we show the result of the matching process which maps the two input tracklets onto the global trajectory/ies of a single or two different persons. Vertical scaling exaggerated for better readability.

model is based on means of covariance descriptors determined by spectral clustering techniques. In the second stage, the result is refined by learning the appearance manifolds of the best matches.

4. TRAJECTORY ANALYSIS

Output of the proposed system is near real-time global trajectory data generated by the described tracking processes within a distributed, dynamic network of cameras. This data is the result of the multi-camera tracker process described above, in which all (local) tracklets (trajectories for individual persons for the entire time they stay within the field of view of a single camera) are integrated into global trajectories in a way that for each individual within the observed area at least a single long trajectory exists (possibly split up into into multiple, disjoint trajectory segments whenever the person is not within the field of view of any camera).

This allows a multitude of analysis techniques such as inspecting individual properties of the observed movement (location, time) in real-time with or without taking into account certain static (e.g. movement through a specific shelf setup in the context of retail store layout planning) or dynamic contexts (e.g. interaction with other persons in order to identify group dynamic in a security application). Each specific domain/application of these analysis techniques requires extensive prior knowledge depending on the complexity of the required type of analysis.

Additionally, the anonymous movement data allows long-term storage and big data analyses for statistical purposes. The data may even pose as context to itself, as security-applications often work with anomaly-detection approaches where a certain amount of prior knowledge is required as reference wrt. what types of movement behavior are common/uncommon in a stable context.

5. CONCLUSION

The system described in this paper has been implemented as prototype system and deployed for proof of concept under real conditions at the entrance hall of the Leibniz University Hannover. It consists of 6 pan-tilt smart cameras which provide MJPG video streams (15 fps) with 4CIF resolution. The concept has been proofed during a three day evaluation phase. The system shows an overall stable performance, particularly wrt. significant

illumination changes over hours, as well as regarding the reduction of false positives by post processing and trajectory merging performed on top of the panorama based detection module. As proof of concept Fig. 4 shows tracking results of single camera tracker and after trajectory merging.

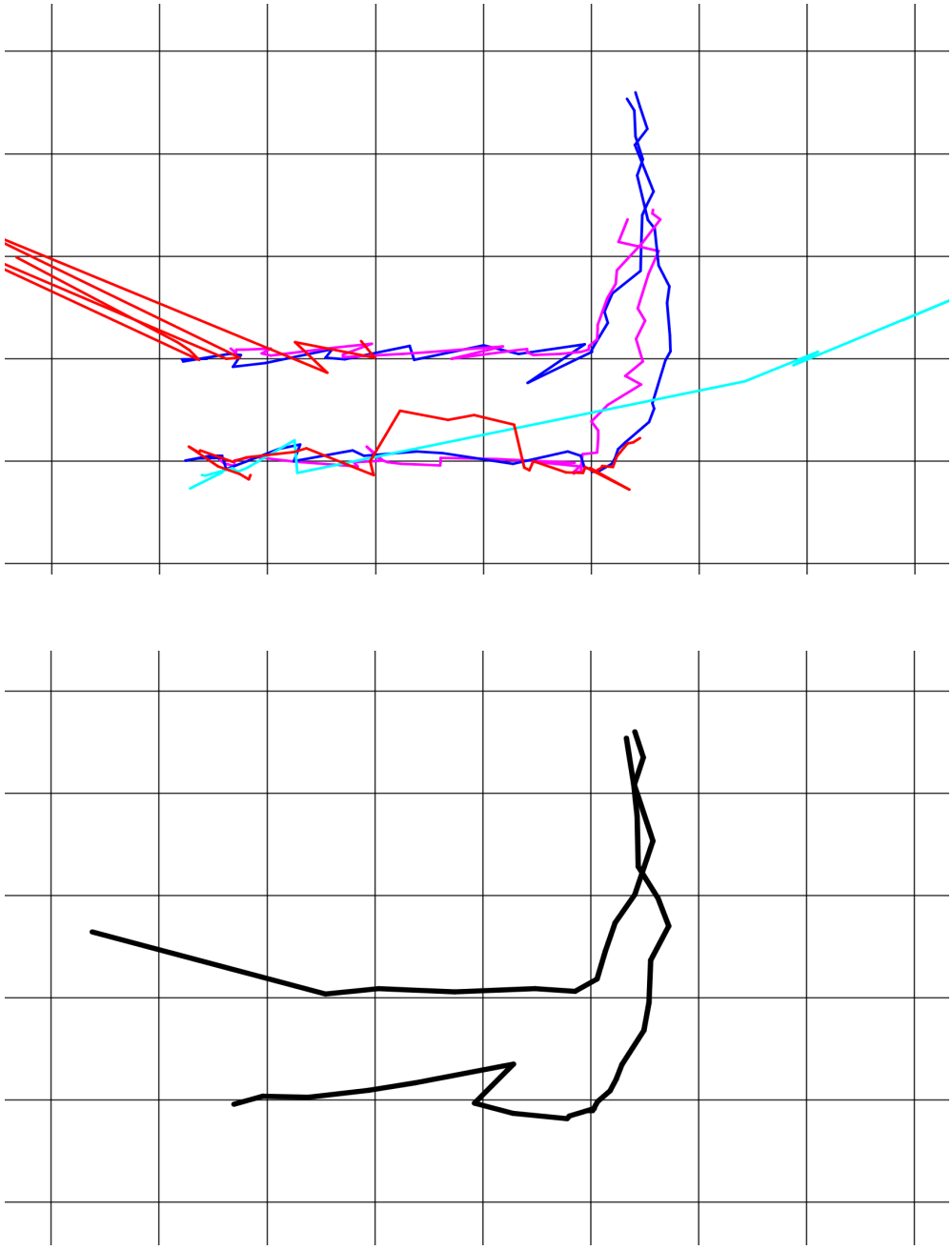


Figure 4: Tracking results of the proposed system: (top) tracks obtained by 4 of the 6 pan-tilt cameras, projected onto the ground plane by geometric transformation given by extrinsic and intrinsic camera parameters and (bottom) results of the trajectory merging approach are shown.

REFERENCES

[1] Collins, R., Lipton, A., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., and Hasegawa, O., "A System for Video Surveillance and Monitoring," tech. rep., Robotics Institute, Pittsburgh, PA (2000).

[2] Cai, Q. and Aggarwal, J. K., "Tracking human motion in structured environments using a distributed-camera system," *IEEE PAMI* **21**(11), 1241–1247 (1999).

[3] Javed, O., Khan, S., Rasheed, Z., and Shah, M., "Camera handoff: tracking in multiple uncalibrated stationary cameras," in [*Proc. Workshop on Human Motion*], 113–118 (2000).

[4] Gupta, A., Mittal, A., and Davis, L. S., "Cost: An approach for camera selection and multi-object inference ordering in dynamic scenes," *Proc. of the ICCV* (2007).

[5] Aghajan, H. and Cavallaro, A., [*Multi-Camera Networks: Principles and Applications*], Elsevier Inc. (May 2009).

[6] Krüger, W., "Robust and efficient map-to-image registration with line segments," *Mach. Vision Appl.* **13**(1), 38–50 (2001).

[7] Monari, E. and Pollok, T., "A real-time image-to-panorama registration approach for background subtraction using pan-tilt-cameras," *Proc. of the Eighth International Conference on Advanced Video and Signal-Based Surveillance AVSS* , 237–242 (2011).

[8] Manzanera, A. and Richefeu, J., "A robust and computationally efficient motion detection algorithm based on $\sigma - \delta$ background estimation," in [*Proc. ICVGIP*], (Dec. 2004).

[9] Herrmann, C., Manger, D., and Metzler, J., "Feature-based localization refinement of players in soccer using plausibility maps," *Proc. of International Conference on Image Processing, Computer Vision, and Pattern Recognition IPCV* **vol. 2** (2011).

[10] Metzler, J., "Two-stage appearance-based re-identification of humans in low-resolution videos," *Proc. of the IEEE International Workshop on Information Forensics and Security WIFS* , 19–24 (2012).