in: Proc. 10th IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS 2013), Krakow, Poland

APFel: The Intelligent Video Analysis and Surveillance System for Assisting Human Operators

Alexander Kolarow[1], Konrad Schenk[1], Markus Eisenbach[1], Michael Dose[2], Michael Brauckmann[2], Klaus Debes[1], Horst-Michael Gross[1*]

[1] Neuroinformatics and Cognitive Robotics Lab
Ilmenau University of Technology
98684 Ilmenau, Germany
alexander.kolarow@tu-ilmenau.de

[2] L-1 Identity Solutions AG
Universitaetsstr. 160
44801 Bochum, Germany
http://www.morphotrust.com/

Abstract

The rising need for security in the last years has led to an increased use of surveillance cameras in both public and private areas. The increasing amount of footage makes it necessary to assist human operators with automated systems to monitor and analyze the video data in reasonable time. In this paper we summarize our work of the past three years in the field of intelligent and automated surveillance. Our proposed system extends the common active monitoring of camera footage into an intelligent automated investigative person-search and walk path reconstruction of a selected person within hours of image data. Our system is evaluated and tested under life-like conditions in real-world surveillance scenarios. Our experiments show that with our system an operator can reconstruct a case in a fraction of time, compared to manually searching the recorded data.

1. Introduction

The gaining interest for security in system relevant infrastructures is commonly met with an increased amount of surveillance cameras. Especially in middle-sized infrastructures, like regional airports, train stations, subways, and shopping malls, this form of security leads only to an illusion of safety, since the larger number of cameras are not met with an equal number of security personnel. The high amount of surveillance footage can often not be managed, which inhibits an efficient crime prevention by active monitoring and slows down the investigation by passive monitoring. A devastating example is the temporary shutdown of a Munich airport terminal in 2010. The terminal was closed for several hours after a passenger hurried through the security gate, and the security staff was not able to track the person in the surveillance footage. Such and similar cases like lost children, abandoned baggage, thefts, and investigating suspicious persons, require an advanced surveillance. To assist security personnel in such cases, we study how a complex intelligent system needs to be designed, so that it can track and reidentify persons in multiple non-overlapping cameras. Using our prototype, an operator can monitor and investigate multiple persons and search through hours of multi-camera footage in a reasonable time-frame.

The remainder of this paper is organized as follows: We summarize related work in Sect. 2. In Sect. 3, we describe our system architecture and the involved submodules. In Sect. 4, we evaluate the video-analysis capabilities of our system and present results from live experiments on a local airport. We end with a conclusion.

2. Related Work

Recently, lots of progress was made in the field of automated video surveillance (AVS). As described in [3], the main goal of AVS is to analyze a large amount of data from several surveillance cameras in real-time and direct the attention of a human supervisor to only the relevant cameras. This task, known as monitoring, is addressed in most current AVS systems, e.g. Knight [17], the VSAM project [1], OBSERVER [4], or NEST [12]. An overview of commercially available systems, mainly focusing on monitoring, is presented in [16]. For a more extensive overview of AVS systems we refer to [18].

A remaining issue of most AVS systems is the inability to investigate the course of events after the detection, termed surveillance video mining [3]. This task includes a cross camera search through all stored footage to detect every occurrence of a person or object of interest. None of the above

*This work has received funding from the German Federal Ministry of Education and Research as part of the APFel project under grant agreement no. 13N10797.

mentioned systems is able to master this challenge, since video analysis in hyper-real-time (multiple times faster than real-time) is a complex task.

The system presented in this paper fills the gap by marking every occurrence of a person of interest in several hours of recorded video within a few minutes (or even within seconds for the last hour) and shows her or his path in a global map. This brings AVS systems to a new level by applying an "after-the-event analysis", an issue that was declared as unsolved in [3].

3. Automated Video Surveillance and After-the-Event Analysis

In this chapter, we describe our system for assisting an operator in the outlined security scenarios. Problem specific vocabulary will be introduced in section "3.1 Example and Definitions". The involved system components are described in section "3.2 Subcomponents". In section 3.3, it is shown, how the results are presented to the operator.

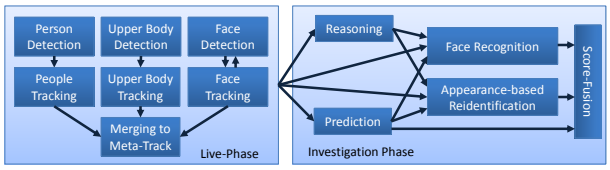

Figure 1. Illustration of the involved subcomponents.

Our human operator assistance system involves a two stage approach (see Fig. 1). In the first phase as much information as possible is extracted in real-time from the live camera footage. This includes people detection and tracking, as well as merging person hypotheses from different methods and cameras into one position hypothesis in global world coordinates. Additionally, features for the recognition modules are pre-computed. All these components run in real-time and store their results in a database for later processing. The second phase is manually triggered by selecting a person. In this phase, different components reconstruct the path of the selected person from first appearance to the point of its current whereabouts. The involved components process the data of the live phase and reconstruct the path of the person through hours of camera footage within a few minutes. The results are then descriptively presented to the operator. Knowing where the selected person was, is and has been in the meantime, helps a human operator to assess the situation much faster than by searching through the video data manually. Additionally, the use of such a system increases the attention of the operator since it actively involves him in the observation in contrast to passively watching hours of data and lots of video streams.

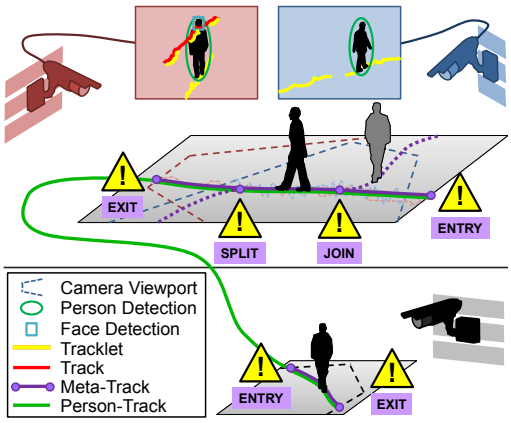

Figure 2. Types of Tracks.

3.1. Example and Definitions

In this section, we describe our system from the human operator's point of view. A common case for security personnel is a lost-child-scenario: In this scenario, a child is accidentally separated from her or his parents. The parents inform the security staff which then tries to find the child in the camera footage. During the incident, all components of the live phase have processed the video data until the current point in time. In the live phase, detectors as e.g. full body, upper body, and face detectors mark the positions of people in the camera. The positions of the people are associated between frames to tracklets by a visual tracking method.

- *A **Tracklet** is as a definite path of one person in a single camera, generated out of person detections of a single detection method and associated by a visual tracking method. To be explicit, tracklets are not associated and continued in critical situations like occlusion.*

Persons close enough to the cameras are detected and tracked by a face detector; these tracklets are combined, if possible, by face reidentification into tracks.

- *A **Track** is defined as univocal path of one person in a single camera, combined of multiple tracklets, assisted by a person reidentification method, like face or appearance-based reidentification.*

All used cameras are calibrated into a global map. Therefore, we transform all tracklets and tracks into global coordinates and combine them to meta-tracks.

- *A **Meta-Track** is a definite path of one person in global map coordinates, generated using multiple tracklets and tracks, associated by proximity, track, and tracklet IDs. In the case of overlapping cameras, those tracklets and tracks are also associated.*

Back to the example, after the parents approach the security personnel, the operator triggers a search back in time on one of the parents. This starts the next phase. The backward search helps to find the point in time, where the child was separated from her or his parents. In this phase, the person recognition modules combine the meta-tracks into a person-track.

- *A **Person-Track** is a definite path of one person through multiple cameras, associated by multiple meta-tracks by reidentification algorithms or a human operator.*

After the child is found back in time in the video data, the operator triggers a new search forward in time to find the current location of the child. The different track types are also illustrated in Fig. 2.

3.2. Subcomponents

Although most subtasks of an automated surveillance system, as for example person detection, tracking, and face recognition are well studied, much adaption and development was required to realize a complete and working system. In this subsection, we outline the components of the live and the investigation phase of our system (Fig. 1). Additionally, we depict the system design concerning data exchange and communication.

3.2.1 Live Phase

Person Detection

Person detection is a well studied field in computer vision. State-of-the-art approaches achieve good results but usually are not real-time capable without the use of specialized hardware.

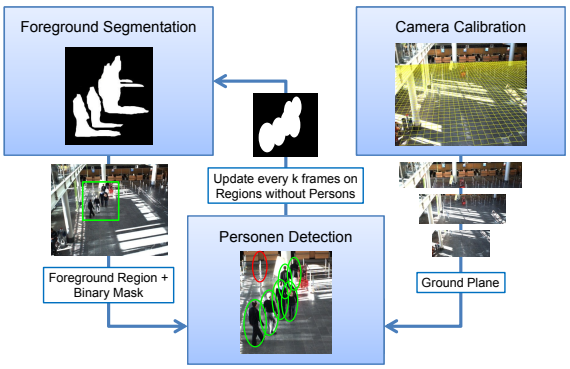

Figure 3. Foreground segmentation and calibrated cameras are used to increase performance and speed of a detection window-based person detector

To increase computational speed and performance of a people detector, we use the setup of Fig. 3, which can be applied for every detection window-based person detector, as for example the HOG person detector [2] or person detection with contour cues [19].

A fast multi gauss foreground segmentation is used to extract regions of interest. The person detector runs only on the foreground regions. The binary mask of foreground pixel is also used to remove false positive detections. Additionally, we use calibrated cameras for further speedup and false detection reduction. With the assumption of a planar floor and known extrinsic parameters of the camera, persons with average height ($1m - 2.2m$) can only appear within certain areas of an image. Only the corresponding areas are used for detection on each layer of the resolution pyramid.

With this setup, the person detection is sped up by an average factor of 112 (factor 8 for segmentation $\times$ factor 14 for the ground plane assumption). We achieve very good results for full and upper body person detection and an average runtime of less then 100 ms per HD-image frame ($1600px \times 1200px$), using contour cues [19], on an Intel Core i7 system. A parallelized implementation for GPU would increase the runtime further.

People Tracking

For ID association within image space, we incorporate a visual tracking algorithm. The common method of track generation using geometrical associations of person detections between consecutive images is not applicable in surveillance scenarios, due to the high risk of ID switches during occlusions. Therefore, we use a template based visual tracker [11]. It generates long continuous high quality tracks in hyper-real-time ($> 100fps$ without the use of parallelized hardware) with occlusion handling and low risk of ID switches. The tracker uses a small set of features, sampled from suitable homogeneous regions to generate a discriminative template. This enables the use of logarithmic search as fast local search strategy. The use of already associated tracks instead of single detections is essential for meta-track creation and person recognition later on.

Face Detection

The key challenge in detecting and tracking multiple small faces in high resolution video data is the real-time constraint. The detector uses a sliding-window approach and compares the content under the window with trained models of faces in a classification step. The models cover a wide range of variations spanned in the direction of illumination, pose and covers biological variations of people from different countries. Due to the large search space (multiple small faces in large images with varying head poses), the detector is the slowest component running with about 2-3 Hz on a consumer PC. To achieve real-time capabilities, we coupled the detector with a fast tracking algorithm. The tracker uses spatio-temporal information for assigning detections into a tracklet. After initiating all tracking threads

(one for each detection), a further full frame detection is started. Using face recognition, the face tracklets are associated into tracks (see Face Recognition component).

Merging Image-based Hypotheses into a Global Meta-Track Representation

Since we use multiple detectors (face, upper body, full body) for each camera, we need to fuse the different hypotheses into a single person hypothesis. Additionally, many surveillance scenarios have overlapping camera views, which makes it necessary to associate person hypotheses between cameras. For this, we transform each person hypothesis into coordinates of a global map. The global map is generated by calibrating each camera using their intrinsic parameters and multiple laser range finders [15].

The global position hypothesis of a person is updated using a Kalman-Filter. To minimize the risk of ID switches in this phase, we use a very conservative approach for data association. Tracklets already associated by image-based tracking obviously do not need further treatment. Hypotheses of different tracklets (e.g by different detectors and cameras) are fused using very strict geometrical restrictions. Whether a hypothesis of a new tracklet is fused to an existing global hypothesis is based on the Mahalanobis distance of both positions using the covariance matrix of the Kalman-Filter and the uncertainty of the hypothesis. The uncertainties of all detectors in a camera are evaluated statistically beforehand and are represented by a covariance matrix. Once a tracklet or track is associated to a global hypothesis, the Kalman-Filter can be updated without further associations.

In spite of the conservative ID association, certain events are likely to cause ID switches or wrong tracks. These events include persons entering or leaving a camera, people walking close to each other, people not being observed due to occlusion, and so on (see Fig. 2). These events need to be handled with great care, since they cannot be resolved later on. ID switches are most likely to occur when two global hypotheses are very close. In this case the meta-track needs to be interrupted for both global hypotheses to assure explicitness. This event is tagged for later processing by reidentification and both hypotheses are fused to a new global non-explicit hypothesis. If two previously joined persons separate, the non-explicit meta-track is interrupted, the event is tagged for reidentification and two new global hypotheses are inserted. In the case of such a split event, explicitness can be restored by reidentification. New and unobserved global hypotheses are also tagged for cross camera recognition. Using reidentification on tagged events ensures that no ID switches occur during critical situations. Merging tracklets with this proposed method into meta-tracks reduces the number of computationally expensive reidentification tasks to only the critical events.

3.2.2 Investigation Phase

Reasoning

To accelerate and improve searches of a person across a camera network, we compute a binary spatio-temporal map that shows all locations - and thus cameras - that a person can reach within a certain time interval starting from a given initial position with an assumed maximum velocity. The computation is based on a wavefront model [10] and uses a map representing the geometrical scene together with the current position as well as a kinematical model of a person. The person and face recognition methods make use of this spatio-temporal knowledge (when a person can appear at which camera device) leading not only to a reduction of the search space but reducing possible false positives at the same time.

Prediction

Since it is necessary to search through all the recordings to find sequences containing the person of interest, it saves time to prioritize the processing sequence of the person hypotheses based on statistics for camera transition and abidance times. Therefore, we utilize a data driven prediction.

The data basis for prediction is encoded into a spatial graph. It is generated in an offline phase by clustering trajectories, obtained by camera [11] and laser-range-finder based people tracking [14, 15]. The mean transition time and variance between neighboring nodes is stored in their connecting edge. The transition probabilities are stored in all nodes for each pair of their connected edges. In the analysis phase, starting with the meta-track of the person of interest, a Monte-Carlo simulation is applied on the graph. The temporal statistics are stored in each node. Afterwards, the stored times are clustered for all nodes in a camera's viewport in order to obtain multimodal temporal intervals of a probable occurrence of the person of interest.

In our system, prediction for one person on an Intel Core i7 took less than 100ms. This simulation reduced the search space for the recognition modules significantly and additionally provided valuable score values.

Appearance-based Reidentification

To track people across multiple non-overlapping cameras, reidentification is needed to connect meta-tracks to person-tracks. Since the point of view can change between cameras, and the visibility of the face cannot be guaranteed, an appearance-based person recognition is needed. The appearance of people's clothes can vary significantly. Therefore, it is important to use a large feature set for reidentification (e.g. texture features [9, 13], color features and histograms [6]), and select discriminative features for a specific person on the fly in the enrollment phase [5]. Using a small subset of well suited features as a template ensures fast matching (12 000 per second). Thus, in the matching

phase this method can easily catch up the spent time for enrollment (about two seconds). For each comparison, a matching score is offered, based on the complement of the probability for a false acceptance, known as inverse logarithmic false acceptance rate score (or $-\log(\text{FAR})$ score). For further details, it is referred to [5].

Face Recognition

The face recognition component is essential for tracking and recognizing a person across a camera network. As a face is a relatively small structure in a video, face recognition imposes some quality requirements on an image, especially with respect to facial resolution. Appropriate camera positions that support the capture of high quality facial images are e.g. near check-in counters, pathways or stairways. To achieve acceptable recognition accuracy, a facial resolution of at least 25 pixels inter eye distance is recommended.

The facial features that are computed during the detection process are used to fit a model to the corresponding image region [7, 8]. At certain points of the facial model, the system computes further features based on Gabor-Wavelet filters which in turn form jets. In order to allow very fast vector comparison statistical signal post-processing is performed on the feature vector. The post-processed feature vector finally gets encoded in a template structure. The computation of such facial features and templates is called template creation. It is executed together with the proprietary detection/tracking component. The recognition process itself operates only on the templates and easily allows for one million comparisons per second.

Score Fusion

To increase the recognition rate, we incorporate the results of multiple modules. For our system architecture, a fusion at score level is preferable. Therefore, the scheme of [5] is applied. We fuse the scores of the prediction module, face- and appearance-based recognition. The score of each module is normalized as inverse logarithmic FAR score. This is done in three steps:

1. Compute a statistic on a benchmark dataset. Select all scores for comparisons of not corresponding persons (FAR statistic)

2. Build a lookup table that holds the accumulated percentage of selected scores s up to the respective lookup entry.

3. For each entry calculate $s^* = -\log(s)$. This avoids floating point imprecision in later calculations.

Due to the normalization of each module's score to the same logarithmic base, the fusion becomes a simple summation of all normalized scores. A remaining issue is the choice of an adequate threshold for the decision if two hypotheses match. This choice should be made problem-specific. We choose a tolerably low threshold, since in our scenario showing a human operator some false positives can be acceptable but missing a true match is not.

3.2.3 Data Exchange and Communication

In the proposed system, a huge amount of data is produced and exchanged between subcomponents. Therefore, a database in combination with a centralized message server is used. The database stores the main amount of data (e.g. tracklets, meta-tracks, features for recognition, etc.). Via short messages the submodules notify each other about new information. Therefore, all submodules can run decentralized. This facilitates an easy expansion of the system and thus guaranties scalability and reliability due to possible redundancy, while data is guaranteed to be saved on a secure centralized system. To further improve reliability, all submodules are designed to cope with lost or unavailable data.

3.3. Visualization

The initial screen for live monitoring is similar to most state-of-the-art systems. The images of the cameras are displayed on multiple screens. The human operator can view the live streams or fast-rewind/forward through the recorded images. For a more detailed view, the operator can enlarge a specific camera to full screen or zoom in on an image. If desired by the operator, various additional information from the live analysis is displayed.

Additionally to this common form of monitoring, the operator can select any person in a camera for further investigation. After a person is selected, a new window opens (see Figure 4). Within this window, all person relevant information is summarized. The current mugshot of the person is displayed in the upper left tile. The upper center tile shows the images of the currently selected camera and time point and is controlled by the time-slider at the bottom. In the upper right tile, the global map of the scenario is displayed with all person positions corresponding to the selected time in the time-slider.

The interactive time-slider is the key element in the person-information window. It is not only used to fast-rewind/forward through the video footage but also displays valuable information of the people-search modules. Frames in which no persons were present or reasoning calculated that the selected person could not have reached that camera in that time, are marked red and can be skipped automatically. Most likely frames for the person to be in are marked with yellow Gaussians by the prediction module. Most important, frames containing the selected person are marked green by the reidentification modules. Additionally, the complete path of the person is displayed in the global map tile (top right). The operator can use this information

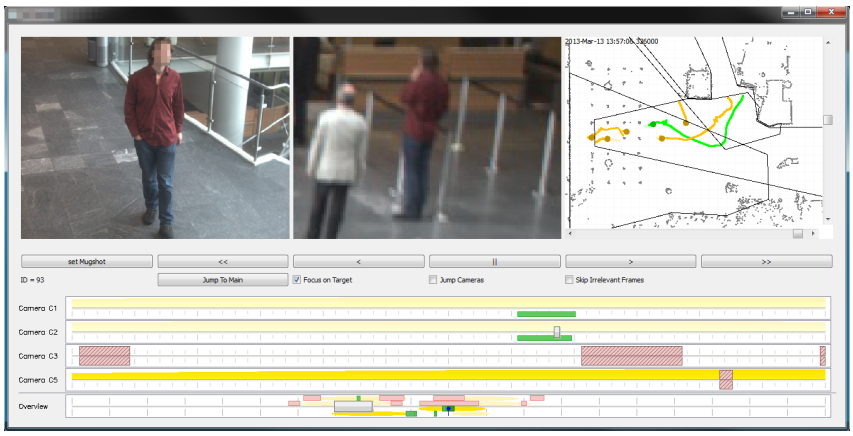

Figure 4. Person specific investigation window. For explanation, see Sec. 3.3.

to jump to the relevant cameras and points in time directly, without manually searching the whole data. Additionally, the upper center tile will jump to the camera with the best view of the selected person and focus on it, while fast playing through the recorded data.

4. Experiments

We evaluate the whole system in two categories: The first deals with the performance of the sub components (Sec. 4.1) and the second additionally incorporates the operator (Sec. 4.2). For the performance of the submodules we refer to the evaluations done in [5, 11, 19].

4.1. Live Evaluation

To test our system, we stage several situations at a local airport[1]. The situations were similar to a lost child scenario and the search for the owner of an abandoned baggage. The system setup consisted of four high definition video cameras ($1600 \times 1200 pixels$), 15 consumer PCs for running the necessary modules and one for the operator to perform the search. The modules needed to cope with changes in lighting, a high passenger volume, non overlapping cameras, occlusions and all other conditions of a real-life scenario.

After the operator started the search, the system was able to provide him with the processed data of all modules within two seconds on average. In one case it took three seconds to analyze the video data but it should be mentioned that this scenario had a length of about 40 minutes whereas all the other had an average duration of ten to fifteen minutes. The enrollment phase of the recognition modules (see Sec. 3.2) took the most time. For live analysis, we did not encounter computational problems even in crowded scenes (more than 50 persons per camera). Nevertheless, we observed that the contour cues detector was only able to process every second frame during high traffic. This had no negative effect since the tracking module still processes the missed frames.

[1]Erfurt-Weimar Airport, 99092 Erfurt, Germany

By extrapolating the computational needs of all modules, we anticipate the real-time capabilities of our system to be thwarted at about 300 persons per camera. The main issue would be the disentanglement of split and join events with the recognition modules. We also encountered some problems in scenes with less traffic, like people standing close together (e.g. at the check-in), or walking in a close group. Due to occlusions, the tracker is not able to follow each person and the appearance-based reidentification is not able to resolve the ID-conflicts as long as no one leaves the group. Even the face recognition is not able to identify the people if the inter-eye distance is below 25 pixels. Fortunately, security relevant scenarios often arise from individual persons, but we intend to address the problem of groups in further work.

Due to data protection regulations we were not allowed to store the video data, which prohibits us from conducting controlled experiments. Nevertheless, the operator was always able to find the person of interest and its complete trajectory in every situation with our system in less than five minutes.

4.2. Saving of Time

In order to evaluate the saving of time provided by the system, we reenacted a theft scenario (scenario 1) on an airfield[2]: Person A removed a radio receiver from a small aircraft standing in a hangar (Fig. 5 A) and put it into a briefcase. He left the hangar, met with person B and handed over the loot (Fig. 5 B). Afterwards they split up and person B tried to leave the airfield with the radio receiver through the main entrance (Fig. 5 C). The operator suspects person B of a theft and wants to reconstruct the course of actions, based only on the 40 minutes of video recordings, comprising almost 30 persons.

The system consisted of four high definition cameras:

[2]Schoenhagen Airport, 14959 Schoenhagen, Germany, European Aviation Security Center (EASC) e.V.

one in the hangar and three outdoors, two with an overlapping view (see Fig. 5).

Furthermore, we reenacted a lost child scenario (scenario 2) on an airport. A family (two parents, three children) entered the airport and turned in their baggage at the check-in. Afterwards, they went to a restaurant, while one child left the group unnoticed in order to go to the toilet. The family noticed the disappearance and informed the security personnel, while the child left the toilet and roamed around in search for his or her parents.

The system also consisted of four high definition cameras: one on the upper floor and three on the lower floor with partial overlap, comprising 80 minutes of video data with more than 30 persons participating.

All video data and the results from live analysis were recorded on harddrives and replayed during the trials. We asked a group of trained operators to reconstruct the course of actions with the aid of our system on one scenario and without the aid on the other one (split 50/50 on both scenarios). They are not only instructed to find the person of interest, but to be able to tell us their walking paths. First, half of the operators were asked to search the data with our system from the point at which person B arrives at the main entrance, or at which he was informed of the lost child respectively. They needed 122 seconds on average to reconstruct the course of actions in scenario 1. In the lost child scenario, evaluated by the other half of the operators, they needed 157 seconds on average. Afterwards both teams used our software to search through the video data of the yet unknown scenario by themselves without the aid of our system. It took them 533 seconds on average to search through the data of scenario 1 and 632 seconds on scenario 2. So the reconstruction of the case was speed up by 4.3 in the theft scenario and by 3.4 in the lost child scenario. This depicts an obvious benefit for operators to use our system for parsing video data for specific persons. Furthermore we suspect that in scenarios with more cameras (20-40 are common) and more people the speed gain would be even greater due to the increased complexity and video footage.

5. Conclusion

Our proposed intelligent surveillance system helps an operator to deal with the increasing amount of camera footage in surveillance. The system extends active surveillance from monitoring to a powerful investigative tool which helps a human operator to evaluate a critical event fast enough to also respond to it immediately. As far as we know, such a surveillance system has never been developed and evaluated under life-like conditions. Our experiments show that using our prototype, a case in a four camera scenario can be solved faster by a factor of 4.3. We expect that in scenarios with more cameras the speedup will be significantly greater.

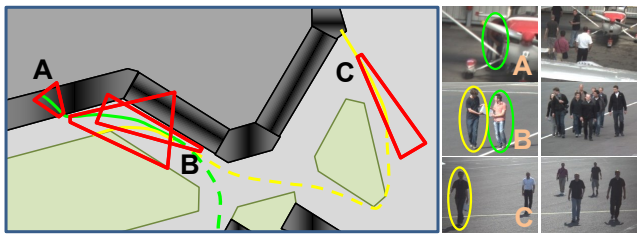

Figure 5. The theft scenario on an airfield. The path of person A is marked green, the path of person B is marked yellow (both walking from left to right) and the viewports of all four cameras are shown as red trapezoids. The key scenes are shown in three images: the removal of the radio receiver (A), the handover (B) and person B at the main entrance (C). The operator had to pick these scenes out of a lot of video footage with several other people walking around (e.g. rightmost images)

References

[1] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson. A system for video surveillance and monitoring. Technical report, Carnegie Mellon University, CMU-RI-TR-00-12, 2000. 1

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 3

[3] A. Dick and M. Brooks. Issues in automated visual surveillance. In *DICTA*, pages 195–204, 2003. 1, 2

[4] D. Duque, H. Santos, and P. Cortez. The observer: An intelligent and automated video surveillance system. In *ICIAR*, pages 898–909, 2006. 1

[5] M. Eisenbach, A. Kolarow, K. Schenk, K. Debes, and H. Gross. View invariant appearance-based person reidentification using fast online feature selection and score level fusion. In *AVSS*, pages 184–190, 2012. 4, 5, 6

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010. 4

[7] S. Gehlen, M. Rinne, and M. Werner. Hierarchical graph-matching. European Patent 01118536.0, 2001. 5

[8] S. Gehlen, M. Rinne, and M. Werner. Hierarchical image model adaptation. US Patent 7,596,276, 2001. 5

[9] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *TSMC*, 3:610–621, 1973. 4

[10] P. Hart, N. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *TSSC*, 4:100–107, 1968. 4

[11] A. Kolarow, M. Brauckmann, M. Eisenbach, K. Schenk, E. Einhorn, K. Debes, and H. Gross. Vision-based hyper-real-time object tracker for robotic applications. In *IROS*, pages 2108–2115, 2012. 3, 4, 6

[12] J. Mossgraber, F. Reinert, and H. Vagts. An architecture for a task-oriented surveillance system: A service- and event-based approach. In *ICONS*, pages 146–151, 2010. 1

[13] P. Salembier and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc. New York, 2002. 4

[14] K. Schenk, M. Eisenbach, A. Kolarow, and H. Gross. Comparison of laser-based person tracking at feet and upper-body height. In *KI*, pages 277–288, 2011. 4

[15] K. Schenk, A. Kolarow, M. Eisenbach, K. Debes, and H. Gross. Automatic calibration of a stationary network of laser range finders by matching movement trajectories. In *IROS*, 2012. 4

[16] M. Sedky, M. Moniri, and C. Chibelushi. Classification of smart video surveillance syst. for commercial applications. In *AVSS*, pages 638–643, 2005. 1

[17] M. Shah, O. Javed, and K. Shafique. Automated visual surveillance in realistic scenarios. *MM*, 14:30–39, 2007. 1

[18] M. Valera and S. Velastin. Intelligent distributed surveillance systems: A review. *PVISP*, 152:192–204, 2005. 1

[19] J. Wu, C. Geyer, and J. Rehe. Real-time human detection using contour cues. In *ICRA*, pages 860–867, 2011. 3, 6