An Intelligent System Architecture for Multi-Camera Human Tracking at Airports

Sebastian Hommel*, Matthias A. Grimm*, and Veit Voges†, Uwe Handmann*, Uwe Weigmann†
*Computer Science Institute
University of Applied Sciences Ruhr West
Germany, Bottrop
E-Mail: (Sebastian.Hommel|Matthias.Grimm|Uwe.Handmann)@hs-ruhrwest.de
†European Aviation Security Center e.V.
Germany, Schönhagen
E-Mail: (voges|weigmann)@easc-ev.org

***Abstract*—In this paper we describe the architecture of an intelligent surveillance system tested at two reference airports. This architecture is developed to support the human operator and enables a multi-camera tracking of suspicious people in case of an alert. The described architecture is based on a network of non-overlapping cameras, each one connected to a self-developed recording tool which provides acquired images to different image processing modules. An efficient preprocessing makes it possible to analyze the data in realtime. The system is able to detect, track and recognize people, but also enables the prediction of where a person will walk to by analyzing possible walking paths.**

***Index Terms*—Airport Security, Decentralized Camera-Network, Fast People-Detection, People-Tracking, Illumination Correction, People-Recognition**

I. Introduction

A central application field of the developed video analytic system is the protection of critical infrastructures, especially at airports. The size and complexity of major airports require an extensive number of cameras for a sufficient surveillance. The fast analysis of the acquired video material from several hundreds of cameras poses a challenge for the security operators. For example, in January of 2010 a part of Munich airport was closed for hours for the reason that the police searched for a man whose laptop had triggered an alert for possible explosives at a security checkpoint. The security wanted to check that laptop again, but the man has left the checkpoint carrying his computer into the terminal. Due to the fact that this man probably just did not know anything about the alert, several hundred people were evacuated and more than 100 flights were affected, whereas the man has never been found. At the same time at Newark airport in New Jersey officials shut down a terminal after a man walked into a secure area without authorization. Videos have confirmed that the man has entered through the exit, but the men's identity could not be determined while dozens of flights were cancelled and thousands of passengers were re-screened. Cases at the airports in Munich or Newark have proven that the recovery of an once detected person over several camera viewpoints is a difficult and tedious task despite the modern video technology that is currently in place. Especially in areas with a high need for security in order to protect human lives it is absolutely essential to react to threats immediately, not least because huge cost arises when terminals are closed and flights are canceled in the case of critical situations.

This research project aims to develop a system that supports the human operator in analyzing video data in order to track and search for people, in case the operator has found a suspicious person and activated an alarm. Although the system, is primarily intended to be used at airports, it is possible to adapt the system to other public areas like train stations or stadiums. A general overview of automated surveillance systems is described in [1]. In this paper we describe the current state of the developed system which is still in progress. Section II gives an overview of the applied hardware architecture whereas the software architecture is described in section III. Section IV outlines the implemented architecture at the General Aviation Airport Schönhagen and finally we will give a conclusion and present our future work in section V.

II. Hardware Architecture

In principal there are two main architectural concepts on how to design the hardware infrastructure: a *centralized* and a *decentralized* video processing which both offer advantages but suffer from different limitations as well. The centralized architecture causes high computational cost on a single processing unit, so this architecture is only usable in small networks [2]. That is why an integration of further cameras is highly limited. In a decentralized system each of the cameras is connected to a dedicated video server and processing unit, so the computational costs incurred are distributed over several processing units [3],[4]. Here, the integration of further cameras is very simple by just adding further processing units. An additional advantage of a decentralized architecture is its reliability in case of a system failure: if one processing unit fails, the overall system is still operable. This is in contrast to a centralized structure. However, the decentralized architecture requires a synchronization of the system time between all video recording units. Therefore, a central time server is necessary.

978-1-4673-5206-2/12/$31.00 ©2012 IEEE

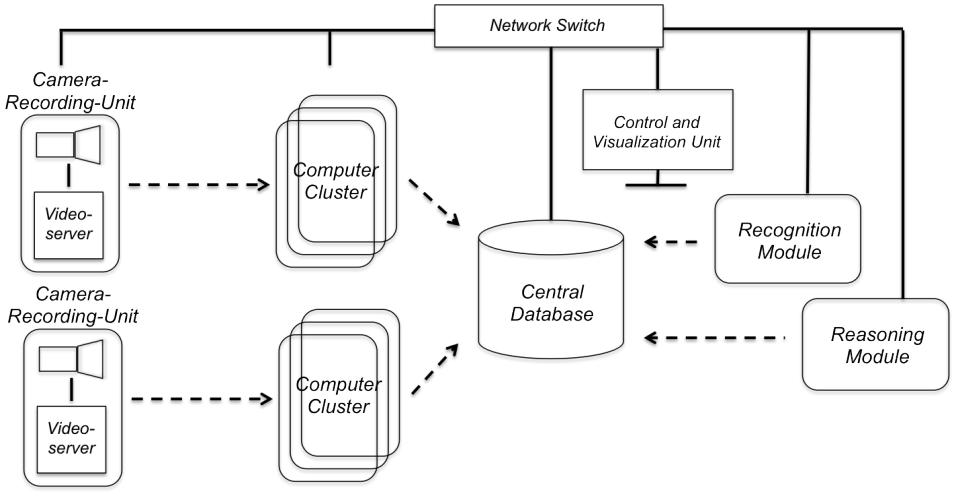

Fig. 1. **Overview of the hardware architecture schematically for two cameras.** The architecture is divided into decentralized camera-based components and centralized components. Decentralized components are the camera-recorder units and the computer clusters for the camera-based data analysis. The database, the control and visualization unit, the module for people recognition as well as the reasoning module are centrally organized.

For the reasons of expandability and load balancing we decided to apply a decentralized architecture for the camera-based analysis. However, a centralized component is necessary which enables the tracking of people across different cameras as well as the location prediction. Hence, the developed system consists of two classes: the decentralized camera-based data analysis, as well as the centralized processing of the extracted information across all cameras. The exchange of meta-data is realized via the central database. The designed hardware architecture is shown in Figure 1.

A. Decentralized Components

The decentralized components analyze the camera-related image data. Therefore, each camera is connected to a video server. A self-developed recording tool acquires the image data and provides several image processing modules with the captured images. The data exchange takes place via a tcp/ip connection. Furthermore, for each camera a cluster of computers (currently three) is used for the live analysis of the acquired images. Here, different image processing modules are analyzing the data, e.g. motion detection, background subtraction, people / head detection and tracking methods.

B. Centralized Components

One part of the centralized system is the control and visualization unit which enables the interaction between the different components of the overall system. This unit is able to visualize different cameras simultaneously, but also allows an interaction with the operator. In the case of an alert the operator selects a suspicious person so that the system knows whom to look for. To track this person in a multi-camera network the recognition module starts to look for this individual across several cameras using the reasoning component. The reasoning component narrows down the search to only those areas which are chronologically and geometrically reasonable. Therefore, a geometric model of the airport is used which reduces the huge amount of data enormously. The location prediction is based on precomputed statistical paths and the individual previously ones which are computed by the recognition module. The hardware requirements of both the reasoning module as well as the location prediction is non-varying even if the number of cameras changes. The communication between the centralized components takes place via the central database.

III. Software Architecture

As the described hardware architecture has shown before, the system is divided into two parts: the decentralized camera-related analysis and the centralized people recognition respectively the location prediction. In this system a considerable amount of sensor data is generated due to the used HD-camera-sensors and the high number of cameras that is necessary to observe an airport. To handle this sensor data, it is necessary to reduce the data volume very fast. This is realized by extracting meta data from live camera sequences during the camera-related analysis. This precomputed information is used to speed up the multi-camera people tracking and the location prediction.

A. Decentralized Components

This analysis is decentralized, in other words each method works separately for each camera.

1) Illumination Correction: For the concrete scenarios which are addressed in this work, indoor and outdoor cameras are needed. However, the illumination differs between different camera locations. Furthermore, the illumination differs all over the day due to the changing position of the sun and dynamically changing whether conditions (sunny/rainy/drifting clouds). By sideward illumination (especially for rooms with big windows), the illumination differs for one camera view in the same time. However, individual features must be similar in different illuminations for the use of multi-camera people tracking. For that reason, a hardware near image enhancement

is calculated before the video stream will be recorded as single JPEG images with a low compression rate. The used image enhancement is based on the camera internal pixel representation. The internal representation of one color channel of one pixel is for modern high-resolution observation cameras higher than 8bit, which is only available for the external image processing. In general, the internal bits are linearly mapped to 8bit for the external image representation. But in our case it is preferable to use a logarithmic mapping function. In this work the preferred function [5] is a gamma correction (equation 1).

$$f(x) = (x/2^n)^\gamma \cdot 255 \quad ; \text{n = number of internal bits} \qquad (1)$$

By the use of this camera internal gamma correction, the image noise does not increase since no sensor information is overrepresented. The gamma γ is estimated by equation 2, where $min(dest)$ is the minimal destination value, $max(dest)$ is the maximal destination value, $min(value)$ is the minimal value of the current image and $max(value)$ is the maximal possible value of the input image. As we mention before, in this work we map to a 8bit image, so $min(dest)$ is set to 1 and $max(dest)$ is set to 256. The current value is set to the mean of the RGB-channels and the overall minimal value is searched for $min(value)$. In our implementation, the input image consist in 12bit for each color channel, so $max(value)$ is set to 4096.

$$\gamma = \frac{log(min(dest)) - log(max(dest))}{log(min(value)) - log(max(value))} \qquad (2)$$

We limit the gamma to 0.63, since the use of a gamma correction with a significantly smaller gamma leads to information loss, since not each of the 8^2 channel values are usable in the external image representation. In the case of video analyzing it is preferable to smooth the gamma temporally.

$$\gamma_{t+1} = \gamma_t + \alpha \times (\gamma - \gamma_t) \quad \alpha \in [0, 1] \qquad (3)$$

This hardware near correction is used in combination with a low brightness threshold for a camera internal automatic exposure time adaptation. So the recorded images are darker with less overexposure whereas the mapping function makes the image brighter and smoother in illumination. So the recorded 8bit image represents some of the previously overexposed image areas just as the well-illuminated areas. With the help of this correction also some areas are represented in the recorded image, which are so dark that they became black in the 8bit image without this illumination correction. The result of this hardware near image enhancement is exemplary shown in Figure 2. The reduction of the threshold for the automatic exposure time adaptation comes to a shorter illumination time which reduces the motion blur.

2) Salient-based People / Face Detection: The next step in the camera-related analysis is a salient-based people detection. First, the foreground will be separated with the help of a background model. This is the first and fastest step to reduce the data volume during the live analysis. Second, an optimized histogram oriented gradient based people detector [6] detects all people in the foreground areas before the model of the

Fig. 2. **Illumination correction.** This figure shows an example result of the presented hardware near illumination correction. The left image is recorded without the correction and the right is recorded with the correction.

background is updated only in those areas, where no people are detected. People detection only on foreground areas speeds up the detector and reduces false detections. For a massively reduction of the effect of learning no or less moving people into the background model, the model is only updated on areas where no people are detected. To handle small illumination changes, a threshold to evaluate the difference between the current image and the background model is used. To handle different illuminated scenes the threshold will be calculated by using the standard variance of the grayscale image. Furthermore, an opening on the binary foreground map is used to reduce the influence of noise. This method is schematically shown in Figure 3.

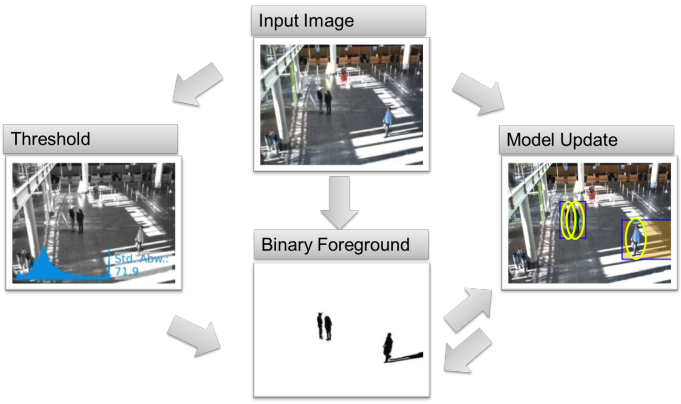

Fig. 3. **Salient-based people detection.** First the threshold to evaluate the difference between input image and background model is calculated and a binary foreground mask is determined. Now, the people detector works at the foreground areas and the background model is updated for all non-person areas.

Equivalent to the people detector, a face detector operates only on foreground areas. Of course, a face is located in the area of a person, so it is imaginable to detect faces only in those areas where a person was detected. In real operations, this could reduce the system performance, since especially in crowded environments full people detections fail while a single face detection works fine.

In our test this method speeds up the people detection about 50 percent in the mean. This method was tested with 5 sequences, which are recorded in a realistic environment at the airport Erfurt-Weimar, Germany. Each sequence lasts one

minute and is recorded with several high resolution cameras at different locations.

3) GPU-based Detection Methods: Especially in security scenarios speed is a very important factor. However, reliable people detection methods like the histogram of oriented gradients algorithm (HOG) are very slow and not suitable for realtime applications. Hence, we decided to use a GPU-based implementation of the HOG algorithm which is based on the implementation of [7]. Our system is equipped with two NVIDIA GeForce GTX 590 video cards, so that 4 GPUs are available for parallelization. The parallel implementation runs over images with a size of 1600x1200 pixels at about 350 ms. Since we need to process images at 10 fps, we revised this approach in such a way, that it is possible to process each incoming image on a different GPU. Therefore, we developed a multi-threading approach which handles the incoming image data. Each time an image is available, it is send to a free GPU which is then blocked until the processing has finished. The processing loop is shown in Figure 4. Using this approach we are able to process 10 fps. In contrast, running the standard HOG algorithm on an Intel Core i7 3,4 GHz CPU, the measured processing time is 3270 ms per frame. Besides the detection of people it is also possible to detect upper body parts and heads by just changing the SVM classifier.

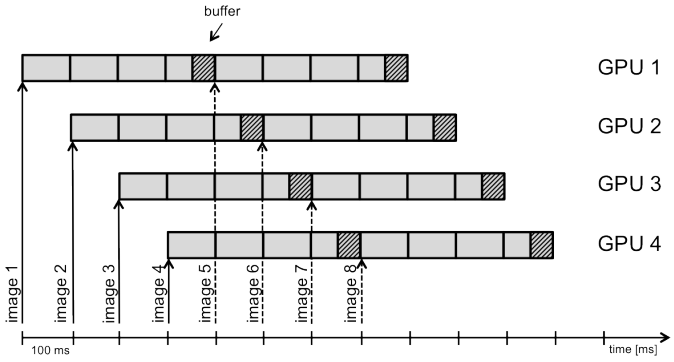

Fig. 4. **Multi-threading approach of the GPU-based HOG implementation.** This approach enables a suitable distribution of incoming images across all available GPUs.

4) Tracking and Feature Extraction: Both the people detections and the face detections are tracked in realtime for one camera view to get first trajectories for the location prediction. These camera-related tracks are also used to find a good representation of a person for the multi-camera people tracking by combining several camera-related views of this person to an individual model by connecting all locations of a person with a similar view, respectively similar features. For the later multi-camera people tracking, first features, like the mean color (hue) and the intensity are extracted from areas of the upper and lower body during this camera-related tracking (Fig. 5). Furthermore, the vertical and horizontal texture rates as well as one histogram of the hue and one for the intensity are extracted only from upper body. Each of this histograms consists in 8 bins.

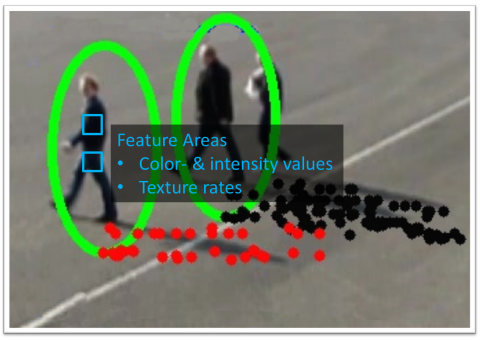

Fig. 5. **Feature extraction during tracking.** First features like the hue, intensity and the texture proportion are extracted during tracking.

This tracking was tested on five sequences with 8 people, which are recorded at the airport Erfurt-Weimar for nearly 50 minutes. In this test no person view is connected to any other, so every view is represented by one track.

To connect the several views, a further tracker is used [8] which is very robust to view changes and partial occultation.

5) Data Fusion: A series of image processing modules is used to analyze the video data, e.g. motion detection, background subtraction or people / head detection methods, which all struggle to perform well under certain conditions. In order to obtain better results and to minimize false detections we fuse the results of the different single modules. Since most of the methods write their results in the form of ROIs into the central database, it is easy to combine the different outcomes. Several approaches exist for data fusion which can be applied on different hierarchical levels [9]. Here, we decided to apply a decision-based fusion. Therefore, in case of a person detection ROI, our fusion module starts to look for intersections with ROIs of other modules. A score is computed based on the intersections. If this score is above a specified threshold, the ROI of the person detection module is considered to be true. However, if there is a person detection that is not part of the foreground (as a result of the background segmentation) and if there is no motion detected in that area, the probability of being a false-positive is very high so that this ROI can be discarded. Another possibility is to fuse the results of person detection and head detection methods in such a way, that the detection of a person is only considered to be true if a head was detected in the upper quarter. There is a number of different suitable possibilities on how to fuse results, like combining upper body with head detections, head with face detections, or even all of them. As a result of the fusion process the rate of false-positives is massively reduced which increases the reliability and leads to a more robust system. This is due to the fact that the decision process has more independent information available. A complete elimination of false detections is not achieved in all scenarios. However, compared to the result of single modules the overall result is considerably better. An example of the fusion process is shown in Figure 6 where the fusion of the results of different image processing modules has eliminated all false-positives.

We tested the performance of the fusion process on different sequences which were recorded on the two reference airports and could achieve a reduction of false-positives of about 97% on average.

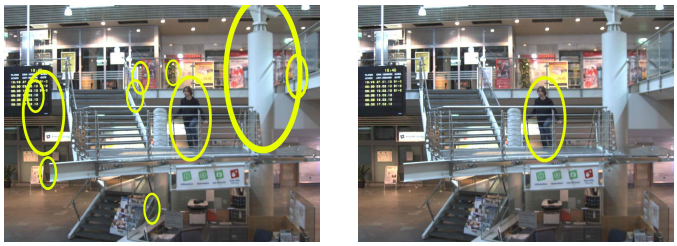

Fig. 6. **Data fusion.** *Left:* Result of a person detection method with a high rate of false-positives. *Right:* Result of the fusion process. Here, the results of motion detection, background segmentation and person detection are fused which, in that case, eliminates all false-positives.

B. Centralized Components

The multi-camera people tracking and the location prediction is centrally realized, since these components interact with video streams and meta data from different cameras. When an alarm is activated, security operators are able to use the presented system to find conspicuous people in the past for a better evaluation of the current situation. Furthermore, it is also possible to find wanted people in the present. To estimate the position of a potential dangerous person in non-observed areas or in the future, the presented system is able to predict this position. This supporting system is based on people recognition which is essential for a multi-camera people tracking with non-overlapping observation areas.

1) Multi-Camera People Tracking: Firstly, an operator marks a detected person and starts the multi-camera people tracking which is based on people recognition. After that, the camera-related tracks of this person will be merged for the current camera view and a person representative will be calculated. In the same time, the temporal and spatial reasoning selects the next possible camera positions and timelines. In this way, this component reduces the search area massively with the help of a geometric model of the operating area. After this, the full body [10] and facial people recognition operates with few selected hypotheses. In this way, the analysis of all recorded image sequences is avoided. For a further speedup of this recognition, the pre-calculated information is used in combination with new extracted information. The full body and the facial people recognition systems generate two hypotheses lists which are separately merged to one final list. Finally, this merged list is presented to the security operator, which confirms the correct hypotheses and decides whether the people recognition stops or will be continued for the next possible camera positions. The computation time of this component is dependent on the number of people tracks for possible camera locations and timelines.

2) Location Prediction: The prediction of the current non-observed people location as well as the future people location is based on a global statistical model and on the individual path model. The global statistical model is generated by analyzing the processes at an airport and mapping them into a concrete geometric model. Furthermore, temporally installed laser range finders are used for an anonymous people tracking in order to generate typical paths [11] and to optimize the global statistical model. To generate the individual path, the multi-camera people tracking system is used. For this location prediction as well as for the reasoning component, a map of the operational area is needed.

IV. Implementation at the General Aviation Airport Schönhagen

For the evaluation and testing of the developed video analytics system a test environment was realized at the General Aviation Airport Schönhagen south of Berlin, Germany. Here, the research work could be applied to a realistic airport infrastructure and the prototype system could be tested under conditions that are not obliged to strict security regulations as is the case at major airports with significant passenger volumes. Furthermore, the GA airports themselves are a targeted application area of this research project, since the developed technology could prove to be a cost-efficient solution for the securing of assets and aircrafts, especially if the security regulations are strengthened.

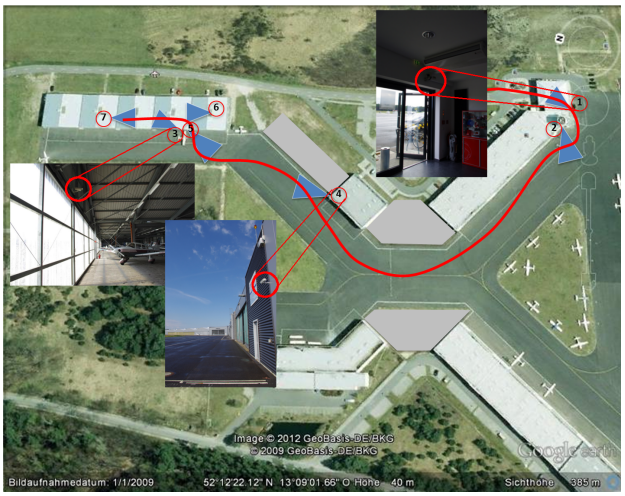

Fig. 7. **Overview of the General Aviation Airport Schönhagen.** The image shows the positions of the mounted cameras. The red line indicates a typical walking path.

As the basis for choosing the different camera positions several scenarios have been constructed to analyze possible routes of persons using the airport infrastructure and to account for versatile lighting conditions and distances to targets. The installed system consists of seven camera positions that allow for a wide surveillance while still being punctual and thus realistic from the user's cost perspective. At three positions the cameras are fix-mounted with weather-proofed housings, while the other four are flexible positions for the testing of different camera adjustments. The cameras are equipped with lenses of different focal lengths. In general, cameras which monitor

a large area are equipped with a short focal length. Those ones which focus on face detection scenarios for example need a longer focal length, which means a smaller field of view but also means a high magnification. In order to evaluate the described system we have developed scenarios with different levels of difficulty. In all scenarios there is one target person, but the number of persons which can be seen in the cameras differs from one person to a group of persons which leads to occlusions. Figure 7 gives an overview of the General Aviation Airport Schönhagen and shows the positions of the mounted cameras as well as the walking paths in our scenarios.

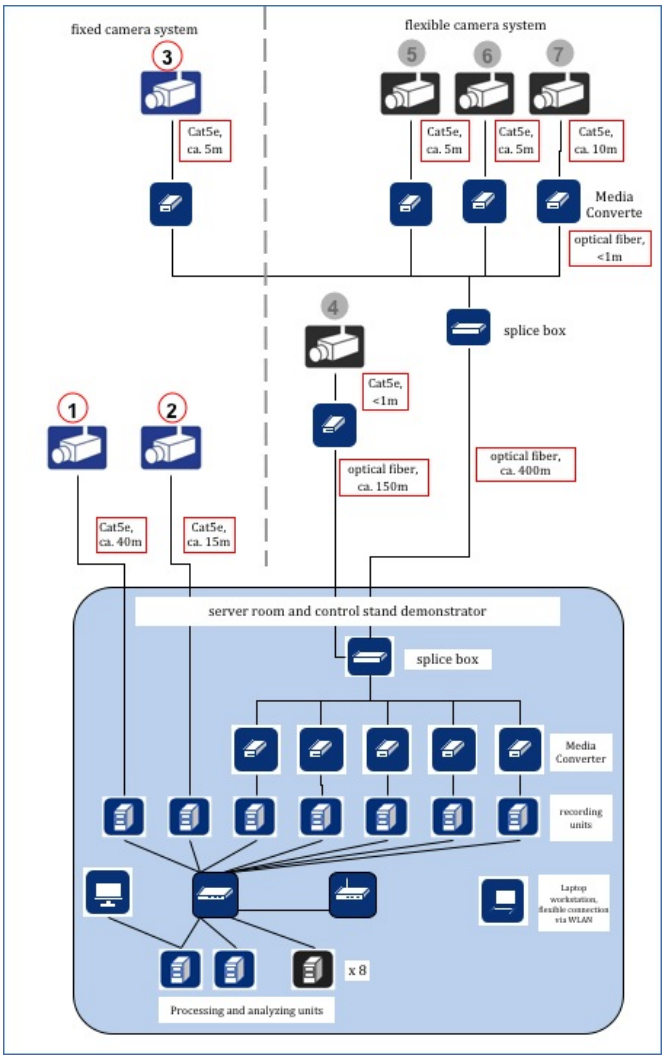

Fig. 8. **Camera system realized at the General Aviation Airport Schönhagen.** The image shows the camera network and how the different units are connected to each other.

All video data is transferred into a server room, which is also working as a security control stand in the demonstration of the system. The cable distance from here to the camera positions varies from 50 meters to about 450 meters. To minimize signal interference and ensure the correct transmission of the large data streams from each camera, the longer distances are covered by using optical fiber cables. The signal to the GigE cameras as well as to the recording computers is then transferred using media converter on both ends. A total of nine computers is installed in the server room, each with standard performance configurations (Intel Pentium Core i7 2.8 GHz, 8GB RAM, 500 GB disk drive, 64 bit Windows 7 Pro). These are connected over a gigabit network switch. For each camera a single computer is dedicated to the recording of its raw video data. While the developed analytics software runs on the remaining computers and further units that are added accordingly. The realized network is shown in Figure 8.

V. Conclusions and Future Work

In this paper we present a system architecture which enables a multi-camera people tracking with non-overlapping observation areas. This architecture allows a scalable number of cameras by adding one image recorder and one decentralized computer cluster for each camera. The system will be further tested and evaluated regarding its usability from the security operator standpoint as well as regarding its acceptance by affected persons, who are using the infrastructure as a pilot, passenger, or airport employee. Furthermore, the methods described in section III will be separately optimized and tested in further works.

Acknowledgment

This work was funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the APFel project under grants 13N10798 and 13N10800.

References

[1] P. Kumar, A. Mittal, and P. Kumar, "Study of robust and intelligent surveillance in visible and multi-modal framework," *Informatica*, vol. 32, pp. 63–77, 2008.

[2] I.-C. Chang, J.-W. Yu, and J.-H. Yang, "Event detection and target tracking based on co-operative multi-camera system," in *ICCE 2009*, jan., pp. 1 –2.

[3] R. Farrell and L. Davis, "Decentralized discovery of camera network topology," in *ICDSC 2008*, sept., pp. 1 –10.

[4] C. Ding, B. Song, A. Morye, J. Farrell, and A. Roy-Chowdhury, "Collaborative sensing in a distributed ptz camera network," *Image Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 3282 –3295, july 2012.

[5] J. Scott and M. Pusateri, "Towards real-time hardware gamma correction for dynamic contrast enhancement," in *AIPR Workshop (W)*, oct. 2009, pp. 1–5.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR 2005*, vol. 1, pp. 886–893.

[7] V. A. Prisacariu and I. Reid, "fasthog - a real-time gpu implementation of hog," University of Oxford, Department of Engineering Science, Tech. Rep. 2310/09.

[8] A. Kolarow, M. Brauckmann, M. Eisenbach, K. Schenk, E. Einhorn, K. Debes, and H.-M. Gross, "Vision-based hyper-real-time object tracker for human-robot interaction," *IROS 2012*.

[9] B. V. Dasarathy, *Decision Fusion*. Los Alamitos: IEEE Computer Society Press, 1994.

[10] M. Eisenbach, A. Kolarow, K. Schenk, K. Debes, and H.-M. Gross, "View invariant appearance-based person reidenti cation using fast online feature selection and score level fusion," *AVSS 2012*, pp. 184–190.

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