Multi-Camera Crowd Monitoring: The SAFEST Approach

Alexandra Danilkina, Géraud Allard, Emmanuel Baccelli, Gabriel Bartl, François Gendry, Oliver Hahm, Gabriel Hege, Ulrich Kriegel, Mark Palkow, Hauke Petersen, et al.

Multi-Camera Crowd Monitoring: The SAFEST Approach

Alexandra Danilkina[1], Géraud Allard[4], Emmanuel Baccelli[2], Gabriel Bartl[3], François Gendry[4], Oliver Hahm[2], Gabriel Hege[6], Ulrich Kriegel[5], Mark Palkow[6], Hauke Petersen[1], Thomas C. Schmidt[7], Agnès Voisard[1,5], Matthias Wählisch[1], Hans Ziegler[5]

[1] *Freie Universität Berlin, Institut für Informatik, Takustrae 9, 14195 Berlin, Germany*

[2] *INRIA Saclay Île-de-France, 91120 Palaiseau (CEDEX), France*

[3] *Research Forum on Public Safety and Security, Carl-Heinrich-Becker-Weg 6-10, 12165 Berlin, Germany*

[4] *SAGEM Défense Sécurité, 100 Avenue de Paris 91300 Massy, France*

[5] *Fraunhofer Institute for Open Communication Systems (FOKUS), Kaiserin Augusta-Allee 31, 10589 Berlin, Germany*

[6] *Daviko GmbH, Berlin, Germany*

[7] *Hamburg University of Applied Sciences, Department of Computer Science, Berliner Tor 7, 20099 Hamburg, Germany*

Contact email: alexandra.danilkina@fu-berlin.de

Keywords: Video Surveillance, People Detection and Counting, Multi-camera

Abstract: This paper presents the current state of people counting approach created for the SAFEST project. A video based surveillance system for monitoring crowd behaviour is developed. The system detects dangerous situations by analysing the dynamics of the crowd density. Therefore we developed a grid-based people counting algorithm which provides density per cell for the global view on the monitored area. Since multiple cameras may observe same parts of the monitored area, the challenge is not only to count people seen by single cameras, but also to merge the views. Therefore we first detect people seen by each camera separately and then sum the results to a global representation. In order to avoid multiple counting of same objects, the output of cameras in the overlapped regions are weighted.

1 INTRODUCTION

In last years we observe the growing number of surveillance systems. Video based techniques, especially, are wide spread in this context. Systems with different functionality and complexity from simple CCTVs to high performance target tracing are developed. In this paper we present an early warning system for people behaviour analysis with intelligent camera nodes among other sensors.

The goal is to provide a video-based crowd monitoring solution for critical infrastructures, where large number of people come together. Dangerous situations may occur when crowd density becomes high. There is a continuous flow of people with mostly no clear movement direction in such public places as check-in halls or boarding counters at airports. Therefore, it is essential to estimate the crowd size and its density of the whole place for further analysis.

We developed a multi-camera system, which monitors people as a crowd. The overall goal of the monitoring system is to provide aggregated crowd information to the command and control for further evaluation. Using several over the network connected cameras, we are able to monitor large areas, so that decisions can be taken based on the global view on the scene.

Furthermore, the privacy is a critical issue in video-based monitoring of public places. For privacy preserving reasons we use infrared cameras, so that no detailed information is captured. Therefore, we preprocess raw data immediately and aggregate crowd data, so that no raw data is stored or sent over the network.

The functionality of the software components we developed and implemented ensures privacy-preserving crowd analysis for the people counting scenario. We provide a synchronised stream of density information for large monitored areas aggregated from distributed sources for the online complex event detection.

1.1 THE SAFEST PROJECT

In SAFEST (The SAFEST project - Social-Area Framework for Early Security Triggers at Airports, 2014) an early warning system for people behaviour analysis with intelligent camera nodes among other sensors is developed. The goal of the project includes crowd monitoring in critical infrastructures, where large number of people come together. The project brings together challenges from various technical areas from sensor hardware and sensor software platform design over communication issues and knowledge fusion for complex event processing. Among technical challenges, SAFEST includes a social sciences part addressing the problem of acceptance for technical solutions by the public.

In this paper, we focus on crowd detection and monitoring part of the SAFEST project. For the crowd behaviour analysis, we have developed an alerting component, which expects a stream of pre-aggregated high-level crowd information for further complex event processing. In this paper, we present image processing steps in order to compute a global view on the scene as input for the high-level analysis. We describe here therefore a scenario, based on crowd density analysis. Intelligent camera nodes extract people information from raw images, which is then transformed to a grid-based density of the whole scene. The aim of the work presented in this paper is to find an efficient way not only to perform data aggregation on distributed sources, but particularly to merge the results avoiding duplicates and inconsistency.

1.2 CONTRIBUTIONS

For the crowd monitoring we developed a people counting solution which can not only provide information about the number of people in the scene for the current situation, but also to visualise the result during the runtime. The developed system is able to detect people looking vertically on the scene and assign the occupied area to each object for further analysis.

The developed multi-step approach is designed for public places and therefore pre-processes raw video streams and aggregates information about detected people without sending detailed eventually private information over the network. The privacy preserving in each of processing steps for people detection, counting and analysis is a key property of the developed solution.

Furthermore, we designed a workflow of software components and underlying layered data aggregation structure. The developed system structure and possibility of automated system configuration allows to employ the developed solution for various real-world set-ups. Once the system is installed, the counting can be performed directly after a short self-configuring step in the initialisation phase. The developed approach is also not depending on the number of cameras in the set-up. It is scalable and can deal with a large number of sources and therefore can provide density information of large monitored areas.

The functionality of developed software components ensures reliable counting from the list of detected people. We designed and implemented a stable geometrical approach for density estimation. An approach for density fusion from multiple camera views is a central data processing point of the system and ensures the merging of distributed people information to one global view on the monitored scene.

The paper is structured as follows: first, we introduce the developed solution for people counting and its contribution. In section 2 we introduce the application scenario and the system set-up for the crowd monitoring solution. In section 4 we describe the principles of the solution and present detailed information about the system functionality. Real world experiments and evaluation of the system are discussed in section 5.

2 CROWD MONITORING TECHNIQUE

We developed a multi-camera surveillance solution targeting indoor monitoring of moving people. The scenario is depicted Figure 1, which show the deployment of the monitoring system, detailed in the following section.

2.1 Application Scenario

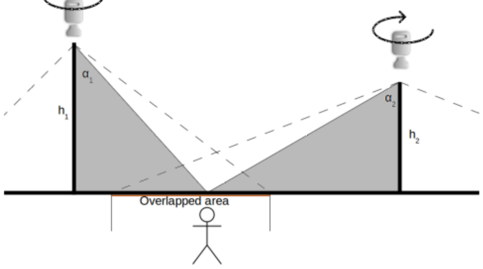


Figure 1: Multi-camera surveillance system deployment.

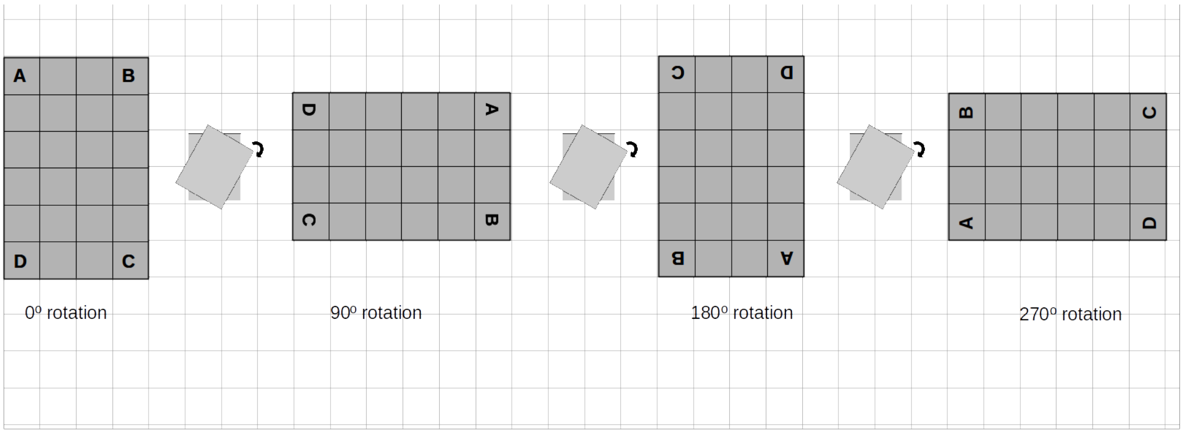
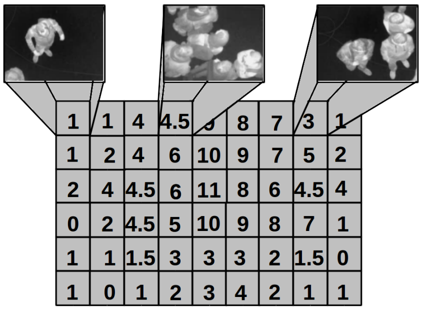


Figure 3: Rotation Explanation.

We assume the reuse of existing infrastructure and install cameras statically on the ceiling of monitored halls. Cameras look vertically on the scene. It is possible to install cameras rotated in the horizontal plane and at different heights (see height h_i and rotation parameters depicted in Fig. 1). However, once the camera is rotated during the set-up, it never moves. Furthermore, cameras may have different view angles (see parameter α_i in Fig. 1). The placement of cameras may lead to overlappings, if more than one camera is seeing the same area. The multi-camera people monitoring solution is developed for two types of cameras: infrared and infrared with depth sensor (Kinect). The system can deal with any set-up with any number of cameras, four possible rotation angles (0°, 45°, 90°, 270°) of cameras in the horizontal plane and any position of cameras in the set-up. The figure 3 explains the rotation process on one exemplary camera by representing the view of the camera by a light grey rectangle placed in the global system of coordinates. The rotation in the global view means rotation of the camera view around its centre of gravity. The crowd monitoring solution is based on video-based computation of the crowd density. We provide the density of a global area for further high-level analysis by computing so called density maps. An example of a density map is presented in figure 2. We divide the monitored area by a virtual global grid of the size of one square meter and estimate the number of people for each cell. Density maps from each camera area are then fused to a global density map. The grid-based data model allows to encode both spatial and temporal feature components in a simple way. We produce one density map per second, so that the dynamics of the crowd density can be evaluated.



1	1	4	4.5	9	8	7	3	1
1	2	4	6	10	9	7	5	2
2	4	4.5	6	11	8	6	4.5	4
0	2	4.5	5	10	9	8	7	1
1	1	1.5	3	3	3	2	1.5	0
1	0	1	2	3	4	2	1	1

Figure 2: Density map visualisation.

2.2 Challenges

The scenario for the people monitoring system comprises two steps. First people for each camera view should be detected and counted. Then the overall grid-based density of the whole scene should be computed based on information from different sources.

For video-based monitoring in public infrastructures, privacy is an essential issue. Since the cameras are placed looking vertically down on the scene, people appear in the scene as "blobs“ can be approximated with circles, so that no personal data leaves camera nodes. This approximation results, however, in an additional challenge. The system should not only detect the "blobs“, but also assign them to single persons. In such areas as boarding control, people may be hectic and push each other and then stand very close to each other and so seen as one "blob“ from above.

Multi-camera people monitoring encounters with further challenge for autonomous monitoring systems, where data from different sources should be handled accurately. Looking on the same area, different positioning of cameras could result in different views on the scene. The challenge is to compute the global view from corresponding sources without being affected by their possibly conflicting input.

The aim of the work presented in this paper is to find an efficient way not only to perform data aggregation on distributed sources, but particularly to merge the results avoiding duplicates and inconsistency.

3 BACKGROUND AND RELATED WORK

In (Wang, 2013) common processing steps solving the main challenges of multi-camera surveillance are analysed. We found again one of the processing steps in realisation of the SAFEST people monitoring system: All cameras should be calibrated before the system can compute the global view on the scene. Since the high-level analysis is based on a grid-based data, a global system of coordinates with axis resolution of one square meter corresponding to a virtual grid was defined. The internal virtual grid of each camera should match axes of the global grid i.e. the camera can be rotated strictly vertically or horizontally in reference to the global coordinate system.

By contrast with many surveillance systems such as presented in (Yang et al., 2003), (Santos and Morimoto, 2011), (Lin et al., 2011), the SAFEST crowd monitoring set-up contains cameras, which look strictly down on the scene. This means, typical challenges for scenes with occluding people cover each other such as perspective estimation, transformation or normalisation are not the focus for the SAFEST solution. Nevertheless, splitting people silhouettes of staying close to each other people is a challenging task.

As presented in (Wang, 2013), in video surveillance systems object re-identification in one camera view as well as through multiple cameras is often required. Works of (Kettnaker and Zabih, 1999) and (Ma et al., 2012) identify people in different camera observations for correct counting. In SAFEST we perform privacy preserving counting and re-identify high-level object – the grids. SAFEST does not focus primarily on people tracking, but counting. SAFEST analyses features, which directly result from infrared camera properties – pixel intensities, which encode the warmth of monitored objects. We do not focus on motion features of the scene as in (Chan et al., 2008) or histogram of oriented gradients (HOG) as in (Ma et al., 2012) and (Zeng and Ma, 2010).

The aim of the work (Teixeira and Savvides, 2007) is to track people, nevertheless the approach of Probabilistic Occupancy Maps is related to the SAFEST density approximation idea. Both estimate the area which is covered by people for further analysis.

The work of (Xia et al., 2011) makes use of the additional depth sensor in Kinect for Xbox 360 cameras (Kinect Camera, 2010) for people detection and further tracking, focuses in contrast of SAFEST privacy preserving approach on the contours of humans though.

4 REALISATION AND ARCHITECTURE

In this section we present the overall solution for distributed privacy-preserving people monitoring system. The figure 4 shows two main principles of the developed solution: the layering approach and the workflow approach. With layers we represent the data processing and aggregation structure, the workflow represents the software structure. We developed four abstraction layers from distributed on-board preprocessing over data aggregation to the data fusion layer, producing input for complex event processing layer. Each of the processing layers may contain several software components from the workflow. The software components are represented with dark grey rectangles. On the lowest abstraction layer, people should be detected. Then single density maps as an abstraction of detected people are computed. The global view on the scene represented by the global density map is a result of data fusion and input for further complex event processing layer, where critical events are detected.

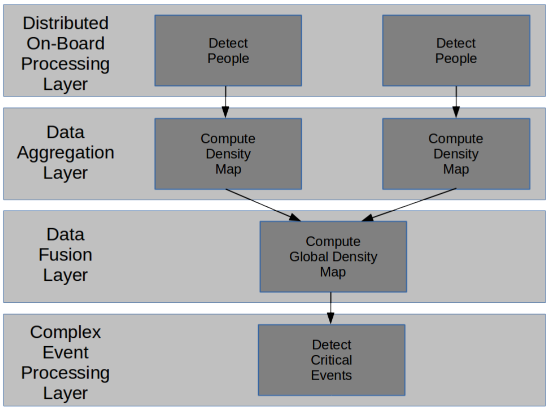


Figure 4: The overview of the crowd monitoring solution.

In order to obtain the final result i.e. a global density map from distributed infrared images, each step of the workflow has to be executed at least ones. Furthermore, software components can be executed in parallel for each abstraction layer, depending on the system set-up. The overall monitoring structure always contains four abstraction layers. Having *n* cameras in the set-up, the on-board processing layer will contain *n* components for people detection, the aggregation layer *n* components for density map computation and the data fusion one component for global density computation.

It is intuitive, that the pre-processing layer is executed on distributed camera nodes and the data fusion and complex event processing on the centralised nodes. However, the data aggregation layer may be implemented depending on available computational power either on the distributed or on centralised nodes.

In this paper we focus on the first three layers of abstraction and workflow steps comprising the video monitoring system, which provides input for the high-level complex event detection of critical situations.

4.1 On-Board People Detection

In this section we present the first step of data processing which is executed directly on the camera nodes. The aim of the software component ”People Detection” is to find people in the stream of infrared images captured by each camera and to provide information about the size of detected persons and their relative positions in the frame.

As mentioned in section 1, privacy is an essential issue in people monitoring systems. Since cameras are looking down on the scene and we are not interested in exact people silhouettes, we approximate people with circles. Position of circles and their radii provide information about the area, occupied by detected people for further density computation. The original frame and the result of the computations are presented in figure 5.

In order to compute the position and the size of circles, two challenges are solved. First pixels representing people should be detected. The second challenge is to aggregate foreground pixels to according circles.

4.1.1 People detection

We implemented two methods for infrared and one method for Kinect cameras for solving the first challenge of people detection. All methods analyse gray levels of each pixel in the raw frame.

The first approach for infrared cameras from (Kaewtrakulpong and Bowden, 2001) is based on the idea of background subtraction. Each input frame is compared to a background model, the difference are foreground pixels. The model of the background is based on the assumption of static background. This approach allows to detect moving people as a foreground and deal with warmth changes in the background. Each pixel of the background is modelled by mixture of three to five Gaussian distributions. The weights in the mixture are proportional to the time, the gray level was observed in the scene. The mixture of distributions are updated during the runtime and the approach is therefore adaptive. We denote this approach with MOG for Mixture of Gauissians from (Kaewtrakulpong and Bowden, 2001).

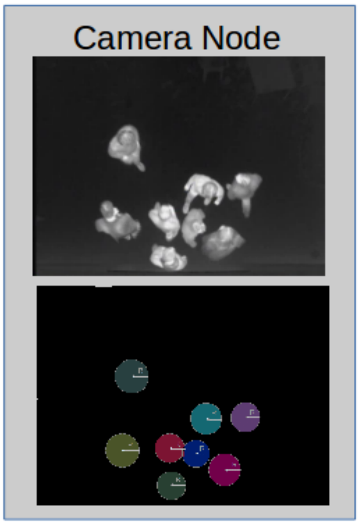


Figure 5: Visualisation of the on-board processing.

The second approach we developed for homogeneous static backgrounds as in presented figure 5. We model the whole scene by one probability distribution of gray levels and represent it as a histogram. People are then represented by peaks in the histogram. By cutting the tails around the peaks, gray levels for foreground pixels are given. We detect peaks in histograms for each frame and therefore the approach is adaptive. We denote this approach with HIST for histogram based approach.

The third approach for Kinect cameras exploits additional information from the built-in depth sensor. We compute the range where people typically may be found depending on the height of cameras in the set-up and apply the HIST algorithm for pixels within this range. Therefore, only peaks within the range are representing people. We denote this approach with K-HIST for histogramm approach for Kinect.

4.1.2 From pixels to circles

The second challenge in the image processing flow is to find single persons in areas representing the whole crowd resulting from the background subtraction. As described in (Baccelli et al., 2014), we first find regions of the foreground, which are directly connected or neighbouring. Connected pixels represent parts people or even whole persons. In order to find parts of persons belonging to one object, we cluster them applying density-based techniques. People who appear in the scene very close to each other may wrongly be detected as one person. Therefore, we split too big clusters exceeding a threshold for cluster size. The threshold depends on the camera height for the current set-up. The last step is to approximate clusters representing people with circles. Therefore, we count pixels, belonging to each cluster as the area of a circle and compute radius from a circle. The centre of gravity for each cluster is a position of a circle within one camera frame.

The output of the software component for people detection on the lowest abstraction level is a list of detected people represented as circles, position of each detected circle and its radius. The stream of this lists is then sent to the next abstraction level "Data Aggregation" from figure 4.

4.2 Data Aggregation for Density Maps

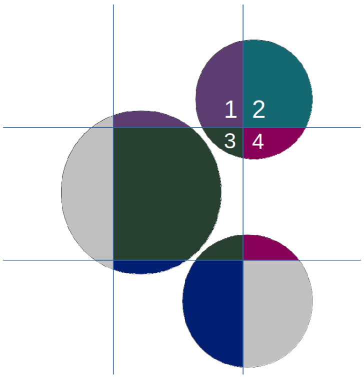


Figure 6: Visualisation of grid crossing detected people.

The software component "Compute Density Map" is designed for deriving stream of densities from stream of people positions on the data aggregation layer for each of camera views separately. The density is represented by the number of people per grid cell in the density map. Simple counting of detected people may be not sufficient in cases, where a person stays directly on the grid separations. The figure 6 shows such cases. People represented by circles and marked with the same colour belong to the same cell and should be accordingly assigned to the cells.

In such cases, we count people proportional to the area, which the circle occupies in the cell. Therefore, a geometrical approach for cutting a circle by a grid was developed. First, for each circle divided by grid, areas belonging to a certain grid cell are identified. An intuition behind the approach is shown for an example circle in figure 6. Areas divided by grid are represented by $i, i = 1,...4$. The areas $A_i, i = 1,..4$ of divided parts of circle and the area occupied by a whole person A are computed in the second step. The person is counted to a corresponding grid cell $i, i = 1,...4$ according proportion $\frac{A_i}{A}, i = 1,..4$ for each cell.

Sixteen possible geometrical cases are worked out and sixteen formulas for parts of circle areas for automated computing of proportions are developed. The output of the software component for density map computations one density map for one list of detected people. The stream of density maps are then sent to the next abstraction layer "Data Fusion" from figure 4.

4.3 Data fusion to a global view

The software component "Compute global density" undertakes a task of constructing a grid-based global view from distributed sources. First, an automatic configuration of the system is ensured. We define camera views and the global view by using the concept of polygons (The Open Geospatial Consortium - OGC Reference Model, 2011): the coordinates of every corner are known starting clockwise with the upper left corner of the unrotated camera. In figure 3 we denoted the corners of the rectangular camera view with "A" to "D".

Rotation angles and positions of cameras in the global grid are sent by the camera nodes to the software component computing the global density. In the initialisation phase, the component computes parameters of the global view: the shape of the global view in a polygon form, number of cameras looking on each grid cell and its position in global system of coordinates.

During the system runtime, streams containing lists with density maps from all the camera views are synchronised – a global view is computed once a second using rounded timestamps for each list. Having the precomputed properties of the global view, the component knows, how many cameras are looking on each cell. For cells with only one input source, the number of people from local density map can be directly accepted. In case of the overlapping, we use following intuition person seen by more than one camera should be counted proportional the number of cameras, so that they are not counted more than once.

The component is able not only to compute the global view for the number of people automatically for any number of rotated cameras, but also to visualise the result in real time. The visualisation was used for evaluation of results as shown in figure 8, which we present in next section.

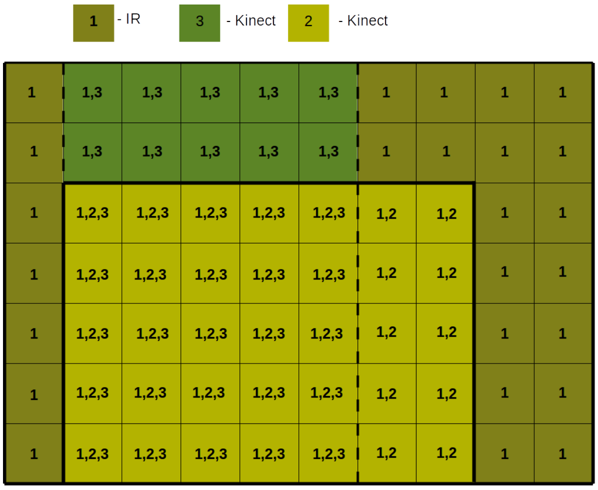
5 EXPERIMENTS AND EVALUATION

5.1 Setu-up and Visualisation

In this section, we present experiments performed for the evaluation of the people counting solution. In order to obtain quantitative results, we set up two series of experiments – real Project partner Flughafen Berlin Brandenburg (FBB), which manages Berlin airports, made possible to install an experiment testbed during the international airshow ILA in Mai 2014. The constructed scenes were captured after, in order to complete the evaluation results with a similar to the real world scenario but a predefined set-up. For both experiments we used the same set of cameras which were placed looking directly on walking people. We used two infrared cameras specially developed for SAFEST by SAGEM (SAGEM Défense Sécurité, 2014) and two Kinect (Kinect Camera, 2010) cameras as a second source of people counting device.

The real-world scenario included two Kinect and one infrared cameras at the same height of 5.7 meters. The placement of the cameras and therefore the overlapping regions result from the available infrastructure: It was possible to place the cameras in the passageway between two exhibition areas, so that mostly small groups or single walking people were observable. The placement of cameras, their rotation and overlappings are shown in figure 7: The infrared camera marked with “1” covers the whole scene and is rotated 180° in the global system of coordinates, the Kinect cameras marked with “2” and “3” and are rotated 270° and 0° respectively.

As described in section 4.3, the component from the data fusion layer configures the global view, establishes the connection to camera sources and draws the visualisation for the experiment. The figure 8 presents the view of people counting system on the set-up. The middle window for the global view shows the counting result for each cell and the overall number of seen people. For better understanding we also marked the borders of each camera view. On the left the counting result of the infrared camera and on the right the counting result for the Kinect camera is presented. Top windows visualise how the data fusion layer sees the incoming counting results from the nodes. The bottom views represent the view of the camera nodes. The views of camera nodes are rotated according to the positioning of cameras.



1	1,3	1,3	1,3	1,3	1,3	1	1	1	1
1	1,3	1,3	1,3	1,3	1,3	1	1	1	1
1	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2	1,2	1	1
1	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2	1,2	1	1
1	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2	1,2	1	1
1	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2	1,2	1	1
1	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2	1,2	1	1

Figure 7: Experiment set-up.

5.2 Counting Resluts

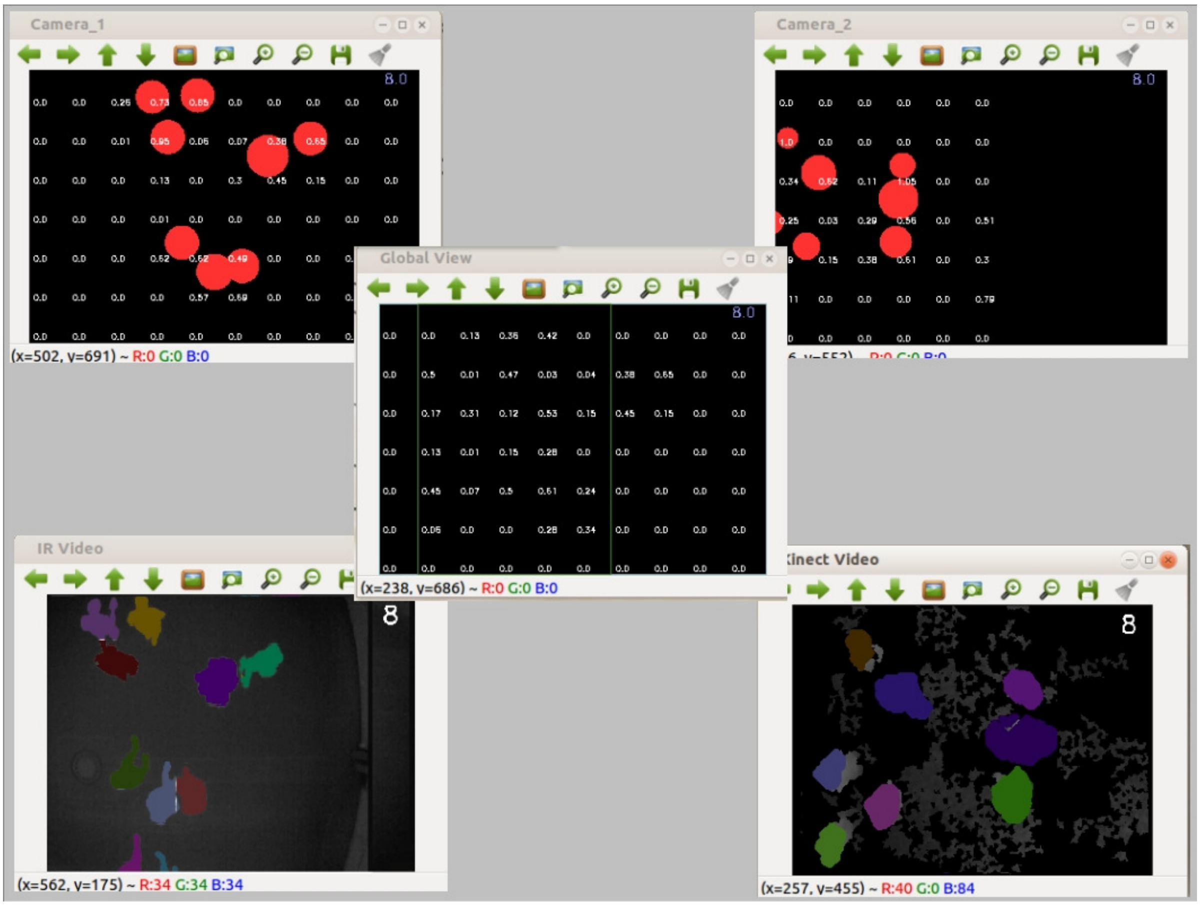
For counting evaluation we recorded video sequences during the real or constructed experiments and hand annotated them for computing relative and absolute errors. Relative error expresses the overall uncertainty of the counting result in percent. The absolute error gives the deviation of the counting result from the ground truth in people. For the evaluation we have chosen randomly distributed timepoints and notated the number of people for the cameras and the merging component compared to the ground truth.

First, we evaluated the counting results for one infrared and one Kinect camera in order to compare different people detection techniques. The table 1 experiment shows the counting results for the histogram-based algorithm HIST, the second for the Mixture of Gaussians background modeling technique MOG.

	Global	Infrared	Kinect
HIST+Kinect	16%/0.42	6%/0.15	21%/0.68
MOG+Kinect	14%/0.47	4%/0.1	19%/0.47

Table 1: Mean relative and absolute errors $\delta x/\Delta x$ for experiment with two cameras.

As we can see, the error difference between two people detection techniques is poor. This results from the scenario. The background was homogeneous, mostly no other objects than people and pushchairs were seen in the scene, the scene was never overcrowded. Nevertheless, we could test and confirm our heuristics for complex cases, where people were



Camera_1

0.0	0.0	0.25	0.73	0.85	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.01	0.95	0.06	0.07	0.38	0.65	0.0	0.0
0.0	0.0	0.0	0.13	0.0	0.3	0.45	0.15	0.0	0.0
0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.52	0.52	0.49	0.0	0.0	0.	
0.0	0.0	0.0	0.0	0.57	0.69	0.0	0.0	0.	
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.	

(x=502, y=691) ~ R:0 G:0 B:0

Camera_2

0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
0.34	0.62	0.11	1.05	0.0	0.0
0.25	0.03	0.29	0.56	0.0	0.51
9	0.15	0.38	0.61	0.0	0.3
11	0.0	0.0	0.0	0.0	0.79
0	0.0	0.0	0.0	0.0	0.0

6, y=552) ~ R:0 G:0 B:0

Global View

0.0	0.0	0.13	0.36	0.42	0.0	0.0	0.0	0.0	0.0
0.0	0.5	0.01	0.47	0.03	0.04	0.38	0.65	0.0	0.0
0.0	0.17	0.31	0.12	0.53	0.15	0.45	0.15	0.0	0.0
0.0	0.13	0.01	0.15	0.28	0.0	0.0	0.0	0.0	0.0
0.0	0.45	0.07	0.5	0.61	0.24	0.0	0.0	0.0	0.0
0.0	0.06	0.0	0.0	0.28	0.34	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

(x=238, y=686) ~ R:0 G:0 B:0

IR Video

(x=562, y=175) ~ R:34 G:34 B:34

Kinect Video

(x=257, y=455) ~ R:40 G:0 B:84

Figure 8: Real-time visualisation of the experiment.

entering the scene very close to each other and were holding hands, so that initially the system were seeing them as a single object.

Furthermore, the table 1 shows the global error. It is less, than the error of the Kinect camera. From this we can follow, that the Kinect and infrared camera were making errors in counting asynchronously. Otherwise, we would not have a global error less than an error of a single source. We developed a weighting function for the global density computation in the Data Fusion Layer, so that counting results from different cameras could be handled differently. Initially, the weighting function was proportional to the camera angle - the more central a person is detected, the bigger is its weight. The intuition behind this concept is decreasing accuracy of counting at the image borders by reason of distortion. The weighting function did not achieved significantly better results. We propose to adapt the weighting function and compute the weights for single camera separately.

The result of Kinect counting shows surprisingly unsatisfactory. Equipped with additional information in comparison to an infrared camera, the counting result was comparable bad. This results from the camera height in this set-up: The depth sensor was not able to reach people from a set-up height of $5.7m$ while the recommended height lies by ca. $4m$ (Kinect Sensor, 2012). Therefore we set-up an additional experiment in order to examine, if the placing of the Kinect camera in a less height leads to better counting results. We placed two Kinect cameras in the backyard of the Freie University Berlin, the possible height was 4.3m and therefore more than recommended though. However the evaluation and the absolute error of 0.46 people and the relative error of counting of 11.3% shows the tendency – the more the height of Kinect camera is closer to the recommended, the more accurate is the counting result.

	Global	Infrared	Kinect 2
MOG+2Kinect	19%/0.2	4%/0.12	14%/0.3

Table 2: Mean relative and absolute errors $\delta x/\Delta x$ for experiment with three cameras.

The second part of the real-world experiment is to evaluate the counting result for three cameras. However, it was not possible to achieve ground truth results for the Kinect 3 from the figure 7. Therefore, we present the error only for the Kinect 2. The global view was computed for all the three cameras though. The table 2 refers the relative and absolute errors for this experiment.

As we can see, the results are comparable to those with two cameras, however the global counting error decreases with the number of cameras. The global counting process relies on two main processing steps: counting people for each camera view and merging the views together. Both steps have influence on the accuracy of the final result. By placing the Kinect camera at the recommended height, a valid solution for the overall goal for distributed multi-camera people counting can be achieved.

6 CONCLUSIONS

In this paper, we presented a multi-camera approach for grid-based crowd density computation developed as a part of the SAFEST project.

We developed a layered data processing and aggregation structure and a software structure. Both and the allow to flexible modelling system employment for various system configurations set up from several distributed sources.

For each of the processing layers we developed and implemented step by step solution for the people counting and perform people detection on the lowest on-board layer, computation of density of the scene on the data aggregation layer, compute single views on the density to a global density view on the data fusion layer and send the result to the critical event detector in the complex event processing layer.

By pre-processing and aggregating the data in each processing step not only the counting accuracy is achieved, people privacy and lightweight data transfer is insured between the software components.

Furthermore, we run experiments in real environment in order to evaluate our solution and confirmed the reasonableness of developed approaches. At the current state of research, we are able to provide a global density information from distributed cameras of two types and visualise the result in real time. Furthermore, the video based multi-step approach was integrated in existing SAFEST structure and is ready for further evaluation.

ACKNOWLEDGEMENTS

The research reported in this paper was sponsored by the Federal Ministry of Education (BMBF) and Research L'Agence nationale de la recherche (ANR) grant, as part of the French-German research project SAFEST.

REFERENCES

Baccelli, E., Bartl, G., Danilkina, A., Ebner, V., Gendry, F., Guettier, C., Hahm, O., Kriegel, U., Hege, G., Palkow, M., Petersen, H., Schmidt, T., Voisard, A., Wählisch, M., and Ziegler, H. (2014). Area & Perimeter Surveillance in SAFEST using Sensors and the Internet of Things. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG2014)*, Troyes, France.

Chan, A., Liang, Z.-S., and Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7.

Kaewtrakulpong, P. and Bowden, R. (2001). An improved adaptive background mixture model for realtime tracking with shadow detection. In *Proceedings of 2nd European Workshop on Advanced Video Based Surveillance Systems*, volume 5308.

Kettnaker, V. and Zabih, R. (1999). Counting people from multiple cameras. In *IEEE International Conference on Multimedia Computing and Systems, ICMCS 1999, Florence, Italy, June 7-11, 1999. Volume II*, pages 267–271.

Kinect Camera (2010). http://www.xbox.com/en-US/xbox-360/accessories/kinect/KinectForXbox360.

Kinect Sensor (2012). http://msdn.microsoft.com/en-us/library/hh438998.aspx.

Lin, T.-Y., Lin, Y.-Y., Weng, M.-F., Wang, Y.-C., Hsu, Y.-F., and Liao, H.-Y. (2011). Cross camera people counting with perspective estimation and occlusion handling. In *Information Forensics and Security (WIFS), 2011 IEEE International Workshop on*, pages 1–6. IEEE.

Ma, H., Zeng, C., and Ling, C. X. (2012). A reliable people counting system via multiple cameras. *ACM Trans. Intell. Syst. Technol.*, 3(2):31:1–31:22.

SAGEM Défense Sécurité (2014). 100 Avenue de Paris 91300 Massy, France.

Santos, T. T. and Morimoto, C. H. (2011). Multiple camera people detection and tracking using support integration. *Pattern Recognition Letters*, 32(1):47 – 55. Image Processing, Computer Vision and Pattern Recognition in Latin America.

Teixeira, T. and Savvides, A. (2007). Lightweight people counting and localizing in indoor spaces using camera sensor nodes. In *Distributed Smart Cameras, 2007. ICDSC '07. First ACM/IEEE International Conference on*, pages 36–43.

The Open Geospatial Consortium - OGC Reference Model (2011). http://www.opengeospatial.org/standards/orm.

The SAFEST project - Social-Area Framework for Early Security Triggers at Airports (2014). http://safest.realmv6.org/.

Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1):3 – 19. Extracting Semantics from Multi-Spectrum Video.

Xia, L., Chen, C.-C., and Aggarwal, J. (2011). Human detection using depth information by kinect. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 15–22.

Yang, D. B., González-Baños, H. H., and Guibas, L. J. (2003). Counting people in crowds with a real-time network of simple image sensors. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 122–129. IEEE.

Zeng, C. and Ma, H. (2010). Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 2069–2072. IEEE.