Critical Situation Monitoring at Large Scale Events from Airborne Video Based Crowd Dynamics Analysis

Alexander Almer, Roland Perko, Helmut Schrom-Feiertag, Thomas Schnabel and Lucas Paletta

Abstract Comprehensive monitoring of movement behaviour and raising dynamics in crowds allow an early detection and prediction of critical situations that may arise at large-scale events. This work presents a video based airborne monitoring system enabling the automated analysis of crowd dynamics and to derive potentially critical situations. The results can be used to prevent critical situations by supporting security staff to control the crowd dynamics early enough. This approach enables preventing upraise of panic behaviour by automated early identification of hazard zones and offering a reliable basis for early intervention by security forces. This approach allows the surveillance and analysis of large scale monitored areas of interest and raising specific alarms at the management and control system in case of potentially critical situations. The integrated modules extend classical mission management by providing essential decision support possibilities for assessing the situation and managing security and emergency crews on site within short time frames.

A. Almer (✉) · R. Perko · T. Schnabel · L. Paletta
Joanneum Research Forschungsgesellschaft mbH, Steyrergasse 17, 8010 Graz, Austria
e-mail: alexander.almer@joanneum.at

R. Perko
e-mail: roland.perko@joanneum.at

T. Schnabel
e-mail: thomas.schnabel@joanneum.at

L. Paletta
e-mail: lucas.paletta@joanneum.at

H. Schrom-Feiertag
AIT Austrian Institute of Technology GmbH, Giefinggasse 2, 1210 Vienna, Austria
e-mail: helmut.schrom-feiertag@ait.ac.at

T. Sarjakoski et al. (eds.), *Geospatial Data in a Changing World*,
Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-319-33783-8_20

Keywords Airborne event monitoring · Automated situation awareness · Video based crowd dynamics analysis · Crowd management · Decision support

1 Introduction

Dramatic examples such as the disasters at Hillsborough football stadium (96 dead and 730 injured persons, 15 April 1989), the Love Parade in Duisburg (21 dead and 541 injured people, 24 July 2010) and recent past the dramatic crowd collapse during the annual Hajj pilgrimage in Mina, Mecca (2.070 pilgrims were suffocated or crushed, 24 September 2015), have shown that safety and security can only be ensured by controlling the movement dynamics of crowds. Critical situations can only be managed properly if they are detected at an early stage, allowing sufficient time to take appropriate countermeasures.

To date, only few studies and empirical data are available that identify dynamic crowd phenomena and derive relevant criteria documenting the behaviour of crowds in critical situations (Helbing et al. 2007). A key aim is therefore to identify the critical parameters and values, to classify situations and to investigate the reliability of criteria based on empirical data. The end users have defined three basic requirements for quickly assessing critical situations and taking appropriate countermeasures:

- Early detection of critical situations before they develop into dangerous situations.
- Clear identification of all intervention options, including information on the locations of all available security and emergency staff as well as available, blocked and open transport routes etc.
- Monitoring the effectiveness of the measures taken to enable appropriate crowd control; ensuring near real-time flow of information including all relevant persons and available information.

Terrestrial video systems are already being widely used for monitoring pedestrian flows. The combined deployment of terrestrial and airborne (airplane and helicopter based, as well as remotely piloted or unmanned aerial vehicle (UAV) based) video systems offers the opportunity to substantially increase the efficiency in detecting critical situations, making fast decisions in due time for coordinated interventions, controlling the measures taken and assessing their effectiveness. The key requirement for joint targeted actions in crisis situations is comprehensive and objective situation awareness on the basis of a situation map. For this purpose, all information required for decision-making and crowd control must be made available to the involved stakeholders. The “Donauinsel Festival 2013”—a highly crowded annual popular music event in the center of Vienna—provided a good opportunity to analyse the processes involved at large-scale events in cooperation with the Federal Police. A plane operator—Diamond Aircraft Industries (DAI)—supported the safety and security management of

the event with a DA 42 MPP[1] that has been equipped with a HDTV sensor as part of a related project—PUKIN (Periodical monitoring of critical infrastructure) which was a research project funded under the Austrian national security research program (see www.kiras.at) provided by the Federal Ministry of Transport, Innovation and Technology (BMVIT). A total of 2.9 million visitors were registered during the 3 days of the festival. The flood of data generated by the event and the time-critical decision-making processes involved clearly showed the necessity for an integrated and automated process to support the decision-makers and security and emergency crews on site and also offer realistic counts or people within defined areas.

2 System Overview and Workflow

Based on the described requirements and the current situation, the aim within the research project was to develop airborne monitoring methods based on video data to enable computer-based analysis of potentially critical movement patterns in crowds as well as an accurate count estimation of people within defined areas. Behaviour monitoring is used to prevent crisis situations by helping security personnel to influence group dynamics early enough so it does not end in critical situation for event attendees. The components developed within the project make it possible to monitor extensive areas and trigger specific alarms in the control centre whenever potentially hazardous situations are detected. All relevant information required for decision-making and crowd control can thus be made available to the stakeholders involved. The following components were developed on the basis of a high-performance multi-sensor video system:

- Video analysis and near real-time geo-processing with the aim of deriving geo-referenced data on crowd density and pedestrian movement.
- Behaviour analysis and simulations for assessing the hazard potential of crowds.
- Intelligent operations control (location based information management).
- Support of field operations through information exchange with mobile units.

Figure 1 gives a rough overview of the included modules and data flows.

The image quality provided by modern airborne and terrestrial video systems enables precise monitoring of crowds and automated analysis of key parameters based on HDTV images taken with appropriate sensor configurations. Figure 2 shows the processing steps required for deriving security relevant parameters from airborne crowd surveillance videos to acquire the current situation providing a better assessment of a current situation for the support of time-critical decision-making processes.

[1] http://www.diamond-sensing.com/index.php?id=da42mppguardian.

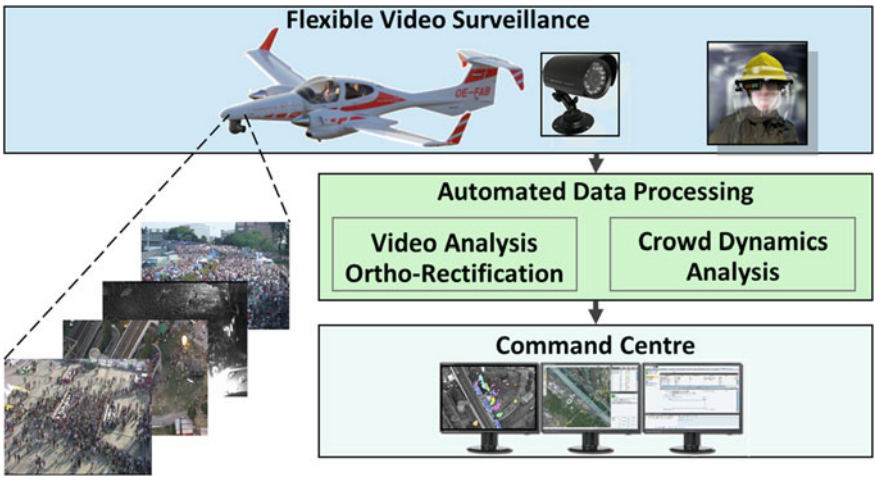

Fig. 1 System components and schematic workflow

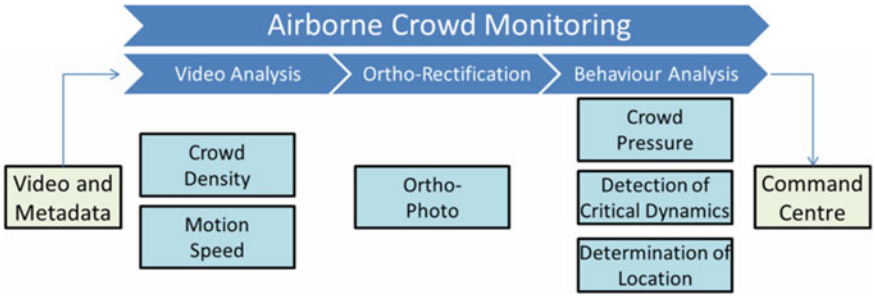

Fig. 2 Workflow for deriving safety and security relevant parameters

3 Airborne Crowd Monitoring

3.1 Video Analysis

The main goal of the video analysis is to fully automatically extract the crowd density, the human count and the crowd motion from a given video stream for each frame. This information is later used to derive the human pressure, which is a crucial parameter to detect critical situations in human crowds. The proposed approach for density estimation and counting is sketched in Fig. 3 and discussed in the next section. The main idea is to extract image features which are then related to the human density by employing machine learning techniques.

The motion is estimated based on the variational description of optical flow in image geometry (Zach et al. 2007). To get a more robust estimate the flow is not

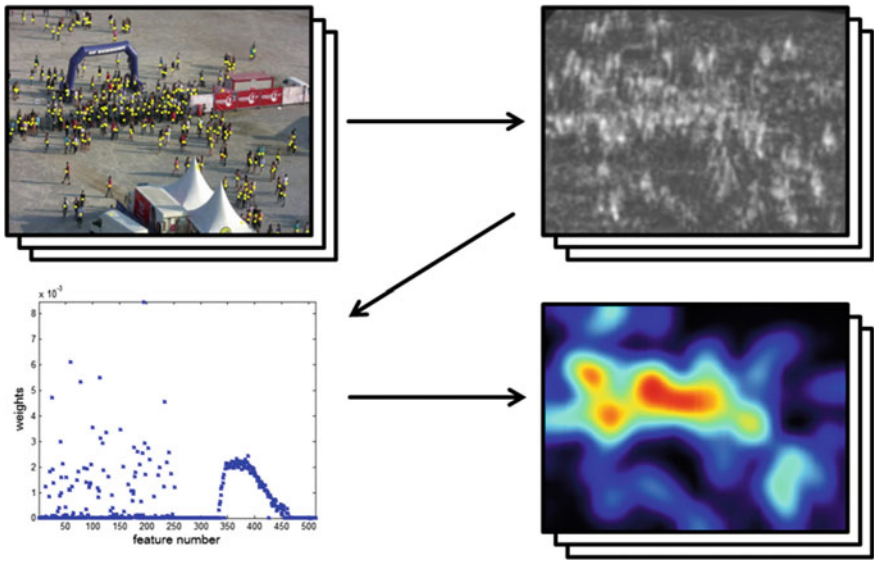

Fig. 3 Proposed workflow for crowd density estimation: An image with annotated humans (*yellow dots*), discretized features (in this specific case the results of an object detector), the learned weights for each feature and the estimated human density function (estimated count equals 250) are shown (Perko et al. 2013)

gathered from two adjacent video frames but from frames with a temporal distance of 10 frames. In addition a given number of those flows are averaged to ensure smooth motion vectors (see Fig. 7). The resulting flow estimate is however in the same temporal sampling rate as the input video, i.e. 25 or 30 frames per second depending on the video camera specification.

3.2 Crowd Density Estimation

3.2.1 Workflow

The presented methodology builds upon preceding work (Perko et al. 2013) which was limited to applications using one specific oblique viewing angle. In such cases pedestrians can be detected based on their silhouettes. However, when the view point changes to more nadir views this method will not produce useful results. Thus, an extension and two novel data sets for testing are presented.

The main idea is to calculate object detection scores from the given images and relate them to the human density by machine learning techniques. As object detector we propose a customized version of the histogram of oriented gradients (HoGs) detector (Dalal and Triggs 2005). The resulting scores are discretized such

that the density estimation method is able to learn a weight for each of the scores. Thus, after learning the density function can be calculated by simple multiplications. In addition, the density estimate is a real density function, meaning that the integral over the density yields the object count (therefore, the integral over a subregion holds the number of objects in this particular region). Example images, the object detection scores and the density estimates are visualized in Fig. 5.

3.2.2 Object Detection

To enable a view invariant person detection we stick to detecting human heads in images, since those are visible in nadir views as well as in side views. Our proposed object detector is based on the construction of a useful descriptor for an image patch. Those descriptors are then used to train a support vector machine (SVM) that is later employed to calculate a confidence score for each location in the image. As basic descriptor we use the well-known HoG descriptor (Dalal and Triggs 2005) which describes an image patch by the occurrence of gradient orientations for a given number of local cells, thus encoding the silhouette of an object. We use the HoG variant reported in Felzenszwalb et al. (2010), since it yields slightly better object detection results while simultaneously having a lower dimensional descriptor compared to the original variant in Dalal and Triggs (2005). For each image patch this implementation results in a vector of dimension $4 + 3*o$ with o being the number of orientations within the gradient histogram. After initial tests we use 9 orientations which results in a 31-dimensional vector for one HoG cell. The size of a HoG cell is set to 15×15 pixels. As one cell would result in a weak descriptor we use 2×2 HoG cells centered on our object and stack those 4 descriptors which finally yield a 124-dimensional feature vector. It can be considered as a rather low-dimensional description especially when compared to the original HoG-based pedestrian descriptor of Dalal and Triggs (2005) with 3780 dimensions. Figure 4 sketches the main concept of our descriptor. Shown is a patch holding a person (left), the gradient magnitude with the spatial arrangement of the 4 HoG cells (center) and one resulting gradient descriptor (right).

For learning we need positive examples extracted from manually labelled objects and negative examples (not holding a person). The positive descriptors are

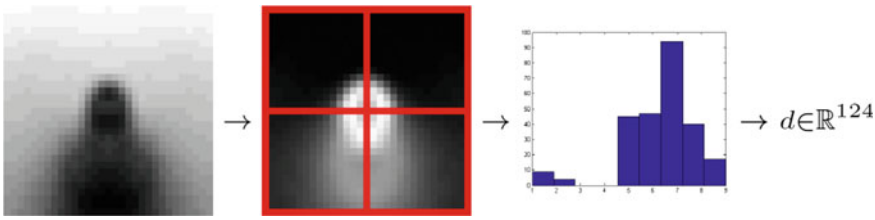

Fig. 4 Sketch of the proposed object descriptor. Four HoGs cells are stacked to gather a 124-dimensional feature vector

calculated for our manually labelled objects, where we are also incorporating a vertically flipped image to double the number of training data. For each image the same number of negative samples is gathered randomly from the image. To avoid that a negative sample also holds a person, a distance transform is calculated from the positive locations. Then a negative sample must have a distance larger than 1 % of the image diagonal (i.e. 18 pixel for an image of size 1440 × 1080).The descriptors of positive and negative samples are employed to train an SVM, where the resulting model is later used for object detection.

3.2.3 Object Counting and Density Estimation

For counting objects and estimating their density we employ the method in Lempitsky and Zisserman (2010). This method takes densely extracted confidences from our detector and learns the density estimate via a regression to a ground truth density. Thus, each pixel has to be described by a feature vector of the following form $f = (0,0,\ldots,0,1,0,\ldots,0)$, which is 1 at the dimension of the corresponding discretized feature and otherwise 0. For density learning our confidences have to be discretized, which is done by setting the minimal value to −2 and the maximal value to +4. These bounds are used to scale the confidences to $[0,255] \in N$. Now, each of the possible 256 values defines a feature vector, as discussed above, which is 1 at the position of the confidence value. Therefore, it yields 256 individual features.

The training itself minimizes the regularized *Maximum Excess over SubArrays (MESA)* distance (cf. Lempitsky and Zisserman 2010) where we use two distinct approaches to solve the resulting linear or quadratic equation system, namely the L_1 and the Tikhonov regularization (i.e. $min_x||Ax - b||$ or $min_{x'}||Ax - b|| + ||(x'\Gamma\ x)/2||$) with $||x \geq 0||$ and Tikhonov matrix Γ being the identity matrix in our case). All details of this methodology are given in Lempitsky and Zisserman (2010). The result is a weight for each of the discretized features and the resulting human density is calculated by multiplying the according weight with the extracted feature value. Thus, for each pixel the density function is given and the sum over all pixels represents the number of objects in the image, i.e. our person count.

Therefore, in the testing phase the discretized features, i.e. our object detection scores, are extracted for each image and multiplied by the learned weight vector, directly resulting in the density estimation per pixel and corresponding person count. It should be noted that this approach introduces virtually no overhead over feature extraction (Lempitsky and Zisserman 2010).

3.2.4 Results of Object Counting and Density Estimation

Test Data Test Data for evaluation of the presented concept videos from three different scenarios were acquired in HD quality. Only individual video frames were used to simulate our envisioned airborne acquisition. Data from other tests showed

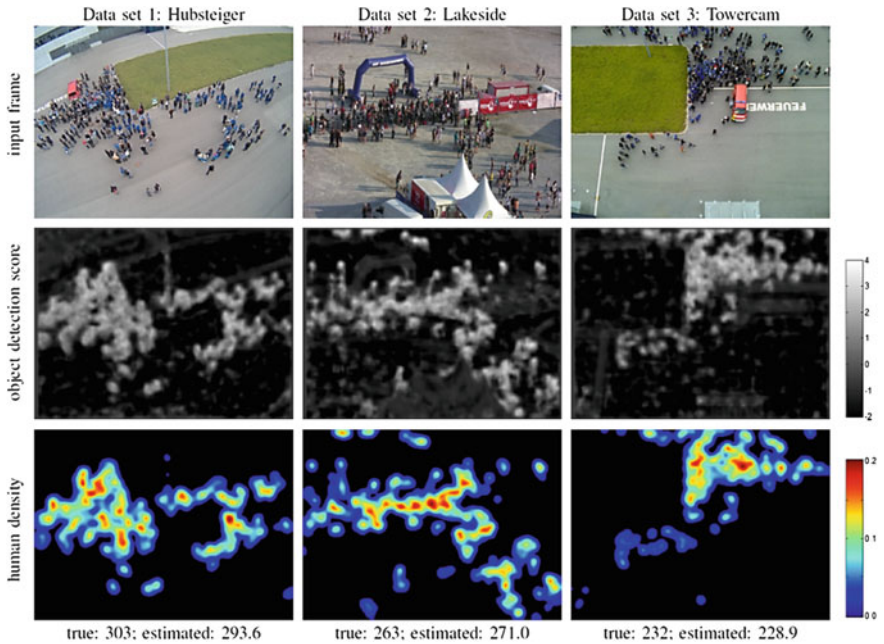

Fig. 5 Exemplary results for the three different data sets. The *top row* shows a representative input video frame for each data set. *Middle row* gives the densely extracted object detection scores which are in the range from −2 to +4. *Bottom row* visualizes the calculated human density functions scaled from 0.0 (*blue*) to 0.2 (*red*) in persons/pixel. The true number of persons in the images and the estimated count are stated below

that the images are analogous to images taken by an aerial platform. Exemplary images are shown in Fig. 5. This figure shows input images, the object detection score and the density estimate. The first scenario, referred as Hubsteiger, originates from a fire drill where we positioned an AXIS P3364 camera on a picker at approximate 25 m above ground. The images of this camera contain fish eye distortion and persons are observed under a slightly oblique look angle. The second one, referred as Lakeside, originates from a music festival in Styria, Austria. A Canon HV30 video camera was mounted on a tower (approximately 30 m above ground). Here the crowd is sensed under a flat look angle of about 14°, such that the whole silhouettes of persons are visible. The third one, referred as Towercam, originates from the same fire drill as Hubsteiger but here a NOKIA Lumia 710 mobile phone was mounted on the top of a building at about 40 m to capture the crowd in nadir direction. Finally, as we want to show the ability of generalization we constructed a combined data set that contains all images from data set 1–3. Even though the presented sequences are not taken from an airborne platform, the images have very similar properties as expected from UAVs or other sensing devices. Therefore, the presented workflow is supposed to yield appropriate results also on airborne imagery. We manually labelled 170 images to get the ground truth person

Table 1 Manually labelled persons in the three data sets together with their statistics

ID	Number of images	Persons				
		Total	Min	Max	Mean	Std
DS1:Hubsteiger	45	11508	15	317	255.7	84.7
DS2:Lakeside	80	22300	249	319	278.8	13.4
DS3:Towercam	45	9468	144	263	210.4	33.4
ALL	170	43276	15	319	254.6	55.1

This information serves as ground truth for training and for testing

counts for training and later for the testing phase (overall more than 43000 persons were annotated, cf. Table 1). From the standard deviation of the people count in Table 1 it can be seen that DS1 Hubsteiger is the most difficult data set, as the number of people changes most dramatically.

Object Detection To evaluate the object detection accuracy we extracted descriptors from positive and negative samples, for each data set and for the combined set (note, that the learning of the combined set involves huge amounts of data, i.e. more than 173000 124-dimensional vectors holding positive and negative samples). Then, we learned Support Vector Machine (SVM) models (Cortes and Vapnik 1995) and calculated the average accuracy by a 5-fold cross validation. For each run 4-folds were used to train the model and 1-fold served for testing. We also compared a linear SVM to a SVM with a radial basis function (RBF) kernel. For the RBF case we also varied two parameters $\gamma \in [0.5, 1, 2]$ and $c \in [2, 4, 8]$ (with γ being a parameter of the RBF kernel function for two samples x_i and x_j with $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ and c being a regularization parameter). While the linear SVM yields accuracies from 93 to 97 %, the RBF SVM performs better with 98.5–99.6 % (best results were achieved with $\gamma = 2$ and $c = 4$). Since the RBF SVM always achieves higher accuracies, this kernel was used for the density estimation later on. The detector for the combined data set gives nice results, with an accuracy close to 99 % using the RBF SVM. Therefore, one detecor will be enough to process all given data sets. For training the final object detector we randomly selected 20 % of positive and negative samples from all data sets and trained a RBF SVM with the parameters stated above.

Object Counting and Density Estimation The accuracy for counting by density estimation of the training and testing process is listed in Table 2. We first used all available images to train the density estimation model. Then we took every second image, then every fourth etc., while the testing of the model was performed on the remaining images. It can be observed that the accuracy of training increases with a lower number of training samples. This makes sense, as the model adapts more and more to the specific samples but loses its ability for generalization (the well-known over-fitting problem). That is why the accuracy of testing is decreasing with a lower number of training samples. Thus, we can learn that about 20 images are sufficient for training the system. We can also observe that the regularization has a small effect on the testing results. Overall, an average error of human counts of about 12 can be

Table 2 Accuracy of density learning and testing

Step	Training			Testing		
	#	L_1	Tikhonov	#	L_1	Tikhonov
1	170	10.3	10.7	0	–	–
2	85	10.0	10.9	85	11.3	11.0
4	43	8.7	9.6	127	11.2	11.3
8	22	7.1	7.4	148	12.2	12.0
16	11	6.2	5.3	159	13.1	13.2
32	6	4.7	3.8	164	13.2	12.9
64	3	2.0	2.4	167	14.9	13.2
128	2	1.2	0.5	168	37.2	23.1

Given are the average errors of the total human count over the training and test images, for two regularization options and different training and test set splits. A count error of 10 represents a relative error of 4 %

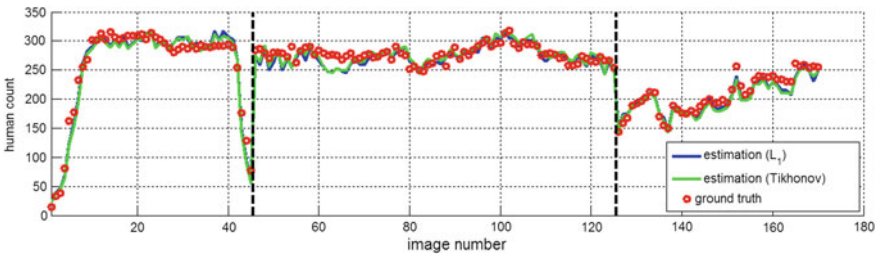

Fig. 6 Person counting: Estimated person count using L_1 regularization (*blue*) and Tikhonov regularization (*green*). The red dots indicate the manually measured ground truth. The *vertical lines* show the transition between the data sets 1, 2 and 3

reached, which correspond to a relative error below 5 %. Figure 5 visualizes some density estimates and Fig. 6 shows the results when using every 4th image for learning. Shown are the estimated human count for the two regularizations given in blue and green colour, together with the manually measured counts shown as red dots. The dashed black lines show the separation between the three data sets. Overall, it can be seen that the estimation is quite close to the ground truth data. Especially for data set 1 *Hubsteiger* our framework is also able to get good estimates when a lower number of people populates the scene (e.g. when people are entering the area in the first few images and when they leave from image number 40–45; cf. Fig. 6).

3.3 Ortho-Rectification

Geo-referencing, also called ortho-rectification, is a standard method in photogrammetry and in remote sensing (cf. e.g. Kraus and Harley 2007) which projects

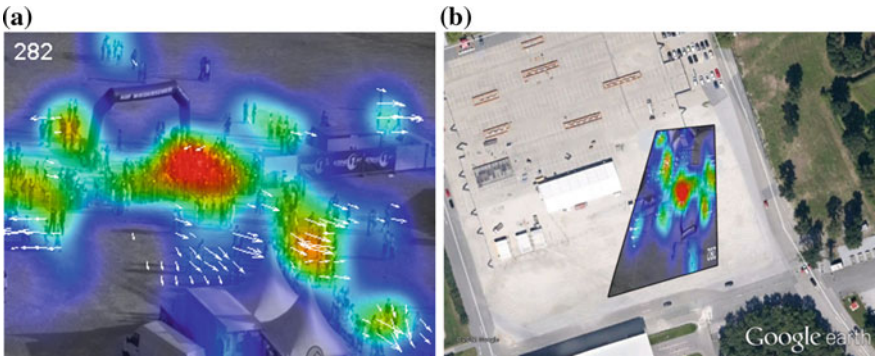

Fig. 7 Geo-referencing of a given image, the human density and motion estimate: **a** input image with superimposed color coded human density function, motion, and estimated number of individuals and **b** the geo-referenced version of (**a**) shown as Google Earth overlay. This conceptual figure is taken from our previous work (Perko et al. 2013)

the image onto the earth's surface with a given map projection. To be able to handle the distortions due to the topography a digital surface model (DSM) is used (global digital surface models like SRTM[2] or ASTER GDEM[3] are freely available). If the terrain is rather flat the DSM can be replaced by the knowledge of the mean terrain height.

All gathered information form the video analysis is geo-referenced and can therefore be visualized and processed in any geographic information system (GIS) system. Figure 7 shows a video frame superimposed with the estimated density and motion and the same information geo-referenced and overlaid in Google Earth. For the ortho-rectification of single images from a video stream, the runtime performance is an important factor. We used an indirect ortho-rectification approach where, based on the camera parameters and the position/orientation of the camera, the position on the ground is calculated. Thereby, we used the raw GPS and IMU (inertial measurement unit) data without post processing steps to enable a near real-time processing. Therefore, GPS and IMU need to deliver positioning estimates with an accuracy which should be sufficiently high for this kind of processing. There are many possibilities for getting the image data and corresponding meta-data from a video sensor, which mainly depends on the used system. For example high end video systems support KLV embedded meta-data in video transport streams like defined within NATO STANAG 4609.[4] For the extraction of single images as well as the synchronized meta-data, we used a commercial product of CarteNav

[2]http://srtm.csi.cgiar.org.

[3]http://gdem.ersdac.jspacesystems.or.jp.

[4]NATO Motion Imagery (MI) STANAG 4609 (Edition 3) http://www.gwg.nga.mil/misb/docs/nato_docs/STANAG_4609_Ed3.pdf

Solutions,[5] which offers this functionality within the product. Depending on the gained format, the data is converted into the internal used meta-data format and used for further geo-data processing. Within this process, using a digital surface/elevation model allows the generation of an ortho-image with an acceptable accuracy for this purpose. The ortho-image itself is projected into a target coordinate system and projection which is appropriate for the area. Thereby we used a metric system as this is easier to use for the following calculation steps.

To keep it simple we define a common map frame for each of our test sites in WGS84 UTM 33 North projection (EPSG Code 32633) since our sites are located in Austria, Europe. Then for each image and for each column/line coordinate the according world coordinate is calculated which are used to rectify the density and motion information.

Density For projecting the density we use a forward transformation and project each density pixel into the common frame. If a pixel gets hit more than one time the values are summed up. This ensures that the sum of the density, i.e. the human count, stays the same in image and world coordinates. Since it happens that some pixels are hit more often than their neighbours due to rounding effects, the whole geo-referenced density is smoothed using a Gaussian kernel.

Motion Rectifying the motion is a bit tricky. In image geometry we cannot differentiate between object motion and camera motion. However, when transforming the reference image coordinate into the common frame using the reference transformation and the corresponding matched image coordinate with the search transformation, absolute world coordinates can be extracted. These two world coordinates define the real object motion independent of the camera movement.

3.4 Crowd Behaviour Analysis

The behaviour of crowds has been studied extensively for many years and researchers have conducted various experimental studies in order to understand human behaviour in different situations. Parameters such as crowd density, speed, flow and crowd pressure, see Helbing et al. (2007), and Steffen and Seyfried (2010) for definitions, are determined either manually (Seyfried et al. 2005) or by means of digital image processing (Johansson et al. 2008; Liu et al. 2009). Many models have been proposed and it was demonstrated that these models can describe the observed self-organizing behaviour like lane formation in crossing flows, intermittent flows at shared bottlenecks, arching at bottlenecks or the transition from laminar to stop-and-go flows (Cristiani et al. 2014). Due to the lack of viable data regarding critical crowd situations, insights of turbulent crowd flows has only been discovered in the last few years.

[5] http://www.cartenav.com.

3.4.1 Turbulent Crowd Flows

Observations of dense crowds showed characteristic motion patterns of mass behaviour like stop-and-go waves or crowd turbulences (Helbing and Johansson 2009). Stop-and-go waves in a dense crowd are first indicators of dangerous overcrowding and can be used for the automatic detection of critical situations that may get out of control and entail disaster. Already Fruin (1993) reported high densities up to 7 persons per square meter, conditions where local crowd density is so high that the all the available space is filled full with human bodies. Under these conditions the individual control is lost and pedestrians become an involuntary part of the mass. In studies of pilgrim flows in Makkah from Helbing and Johansson (2009), stop-and-go waves have been observed even in areas without any obvious bottlenecks. They occur when the pedestrian density reaches a high level such that unobstructed pedestrian flow is inhibited. While the transition from laminar to stop-and-go flow was already well understood the insights of the subsequent transition into turbulent crowd dynamics were first revealed in Helbing et al. (2007). Therein variables were identified which are useful for an early warning of critical crowd conditions. Turbulent crowd flow occurs in situations of extremely high densities and is characterized by movements into all possible directions. It is caused by people who move involuntarily and induce sudden movements of other people nearby. As a consequence, people are pushed around and fall down. They are trampled down and, moreover, they turn into obstacles for others leading to more stumbling people.

3.4.2 "Pressure" in the Crowd

In contrast to purely density-based assessments, critical situations like shock waves and crowd turbulences are characterized by high variance of motion magnitudes under high densities. Helbing et al. (2007) introduced the crucial parameter to detect critical situations in human crowds as the so called local "pressure" $P(\vec{r})$ defined by

$$P(\vec{r}) = \rho(\vec{r}) \mathrm{Var}_{\vec{r}}\left(\vec{V}\right) \tag{1}$$

with the local pedestrian density $\rho(\vec{r})$ times the local velocity variance $\mathrm{Var}_{\vec{r}}(\vec{V})$ of the velocities $\vec{V}$ at the location $\vec{r} = (x, y)$. The local density is calculated by the average circular region of radius R at a given location in conjunction with a Gaussian distance-dependent weight function centred at the location, see Helbing et al. (2007) for details. For instance, the criticality in the crowd depends on the local conditions and therefore close measurements within a radius of $R = 1$ m were used.

3.4.3 Method Employed

As described before, to evaluate critical conditions from a video stream, the crowd density and crowd motion has to be extracted for each frame through video analysis. The novelty of this approach was the use of airborne video instead of terrestrial video to provide an accurate crowd density and motion estimate for automated crowd behaviour analysis.

Accordingly, the results of the video analysis with pedestrian density and motion estimates for each video frame are transformed into a world-coordinate system using ortho-rectification and aggregated cell-by-cell in a two dimensional Cartesian grid with a cell size of 1 × 1 m. For each cell c_{ij} the values of pedestrian density, velocity and pressure are stored in a two-dimensional array. In Fig. 8a, the video frame transformed into world-coordinates is shown. The contour plot in Fig. 8b illustrates the values for density and the green arrows represent the pedestrian velocity in the form of a vector field. In the shown example the grid has an overall dimension of 58 × 92 m and the transformation in the world-coordinate system enables the use of physical units directly, e.g. number of persons per square meter and velocity in meters per second.

To calculate the local velocity variance, the velocities $\vec{V}(c_{ij}, t)$ of the current and the eight directly adjacent cells (Moore neighbourhood) are used instead averaging over a circular region with a radius R, covering an area of 9 m^2. For each cell c_{ij} of the grid the velocity variance is

$$\mathrm{Var}_{c_{ij}}\left(\vec{V}\right) = \langle \left[\vec{V}(c_{ij}, t) - \vec{U}(c_{ij})\right]^2 \rangle_t \tag{2}$$

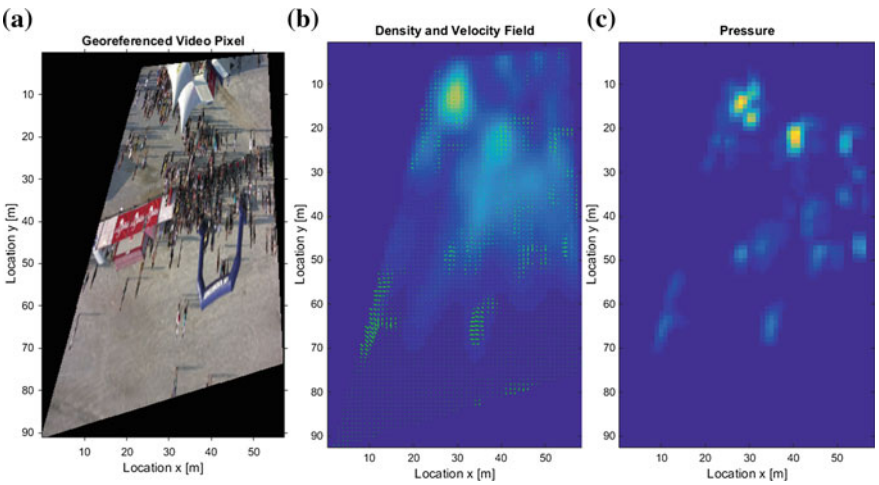

Fig. 8 Geo-referenced video frame (**a**), density with velocity field (**b**) and the computed "pressure" (**c**)

with $\vec{U}(c_{ij})$ as mean velocity in the Moore neighbourhood. By adapting Eq. 1 to the grid structure, the pressure for each cell c_{ij} is calculated with

$$P(c_{ij}) = \rho(c_{ij}) \mathrm{Var}_{c_{ij}}(\vec{V}) \tag{3}$$

whereas $\rho(c_{ij})$ denotes the corresponding cell density. The contour plot in Fig. 8c displays the pressure in the cells (blue colour corresponds to low values, orange and yellow reflect higher values), where regions with higher values indicate higher crowd pressure.

Due to the lack of video data from critical situations, the hot spots shown in Fig. 8c reflect the highest pressure in the current video frame with an uncritical situation only. Here, the highest pressure values reaches $0.002/s^2$ and were only a tenth of the threshold for dangerous situations as mentioned in Helbing et al. (2007) with $0.02/s^2$.

If a critical threshold for the crowd pressure will be exceeded an alert is sent to the command center, including the severity of the situation with the values normal, dense or critical depending on crowd density and pressure, the world-coordinates and a video frame of the video sequence that triggered the alarm.

4 Data Management

The results from the data processing and analysis modules are stored within a geo-oriented data archive. It allows geo-based as well as time based access to the available data and so provide all available data to command and control systems in the command room as well as to mobile units in the field. Thereby the command room provides access to all information required to support the safety and security management of large-scale events. The key aim is to generate a comprehensive current situation map giving a clear overview of the situation while also ensuring a high level of detail.

The main focus at the control room contains three key tasks: (1) Situation awareness, (2) decision support and (3) command & control. Being able to provide accurate information at the needed level and for the relevant target group is an important factor to support decision making processes. Therefore an application was developed to provide the information to the end user in an easy and intuitive way while displaying the position of sensors, infrastructure objects as well as mobile units combined with related sensor data. Also live video streams can be displayed by connecting the client to the streaming server on the sensor. The geo-referenced data is displayed as layers on a base map. These layers include current aerial images, the hazardous areas automatically calculated from the data as well as event-specific maps (e.g. stages, etc.). Access to the data archive allows the integration of historic data and so also allows the analysis how situations arose.

The creation of a common operational picture focused on providing fast and continuous situation awareness, role based data distribution (including commanders,

field commanders and field staff) and a concentration on a high usability for an effective support within crisis situation. Next to spreading information to mobile units, the data distribution also include interfaces to already existing command and control systems (e.g. see Ruatti Commander[6]) or local GIS systems.

5 System Demonstration

The number of high-performance video systems for civilian applications is growing rapidly in many European countries. The stakeholders involved and representatives of the air police (Federal Ministry of the Interior) have confirmed in personal communications that automated processing of video data and crowd behaviour analysis are essential in deploying these expensive multi-sensor video systems for purposes of safety and security management. The approach of multi-sensor data acquisition, geo-oriented data processing, automated behaviour analysis and target group specific representation in an integrated control room has clearly confirmed this potential. The system components were tested in two events. For organisational and legal reasons, only a limited amount of airborne video data was available for the development and evaluation of the system components. Additional video data were recorded from a tower during the Lakeside Festival in Austria in 2011 to simulate the airborne sensor configuration. In June 2012, project partners recorded extensive video data using a FLIR Star Safire 380-HD video system[7] during an overflight of the Donauinsel Festival.

6 Conclusions and Outlook

The presented system for decision support on large scale events yields promising results and helps to detect critical situations in near-real time. However, the limited availability of the video sensor system for legal and organisational reasons turned out to be a key problem in the project. This made it necessary to test different airborne video systems and to simulate airborne imaging situations using terrestrial video systems. Licence problems on the part of the supplier prevented access to the camera control software and the geo-sensor data (GPS, IMU). This reduced the amount of data available for development and testing in the fields of geo-processing, video and behaviour analysis and real-time data processing. The results therefore do not provide reliable information about the performance of the planned video-based crowd monitoring system but show results from the performed tests and different video systems. Even though the exemplary datasets hold only a

[6]Ruatti Systems GmbH, http://www.ruatti-systems.de/en.

[7]www.flir.com/surveillance/display/?id=64505.

small number of people (around 300) and the crowd density is far from reaching panic densities, the presented approach is scalable provided that the image resolution in cm/pixel stays similar. In case of denser crowds the best image acquisition scenario will be a bird-eye view (close to DS3) as in this case human head are still visible and less occlusions occur in comparison to oblique look angles. In addition, a given larger area of interest could be observed by either using a higher-resolution camera with appropriate lens to assure the envisaged ground sampling distance or by employing the movement of the airborne vehicle to cover the area with time-delayed acquisitions. A challenge in the future is to get open interfaces to the sensory of such high end camera systems. These are not only of technical nature because, they principally existing but are not open accessible and not implemented on the different systems. The optimally suited imaging sensor obviously depends on the specific application. When the focus is on observing large scale events it would be beneficial to use a very high resolution camera that is still able to capture at least 3 frames per second, e.g. a Prosilica GT6600 with 29 MPixel and 4 fps. This frame rate is sufficient for motion estimation. When the area of interest to be observed is smaller any consumer grade full-HD video camera will fulfil the requirements. In a succeeding research project, this part as well as the correct synchronisation between individual video frames and the correct metadata from the GPS/IMU sensory will be focused on.

The results obtained from the video analysing component allow to conclude that the motion of crowds can in principle be well monitored and analysed at large scale events, and that the estimates of crowd densities which are crucial for an alerting system are sufficiently accurate to consider the application at future events. This first proof of concept requires future work to investigate the performance of the proposed methodology more deeply, i.e., using additional video sequences and to estimate the potential to adjust it to various illumination and imaging situations. Controlling the image gathering and tailoring it to the specific situation will additionally and substantially improve the results of the proposed image analysis. Oblique imaging also poses serious challenges, as persons may be partly hidden and geo-referencing in uneven terrain (bridges, installations etc.) requires high-precision surface models. These questions could not be dealt with in detail as the available resources were required for testing the different sensor systems. The final challenge will be to reduce analysis time by developing algorithms for near real-time analysis.

Based on the performed demonstrations, feedback from involved security staff as well as end users (police) have shown that the presented system will be a huge support for time critical decision making processes of the security staff in charge. Getting objective and reliable information about crowds and their dynamics allows better responses to critical situations and can so drastically enhance the safety of attending people.

Acknowledgments This work has been partially funded by the Ministry of Austria for Transport, Innovation and Technology within the Austrian Security Research Programme KIRAS: Project 845479: “MONITOR: Near real-time multisensor monitoring and short-term forecasts to support the safety management at mass events”.

References

Cortes C, Vapnik V (1995) Support-vector network. Mach Learn 20:1–25
Cristiani E, Piccoli B, Tosin A (2014) Multiscale modeling of pedestrian dynamics, vol 12. Springer
Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the conference on computer vision pattern recognition, vol 2, pp 886–893
Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part based models. IEEE PAMI 32(9):1627–1645
Fruin JJ (1993) The causes and prevention of crowd disasters. Engineering for crowd safety, pp 99–108
Helbing D, Johansson A, Al-Abideen HZ (2007) The dynamics of crowd disasters: an empirical study. Phys Rev E 75:046109
Helbing D, Johansson A (2009) Pedestrian, crowd and evacuation dynamics, encyclopedia of complexity and systems science, vol 16
Johansson A, Helbing D, Al-Abideen HZ, Al-Bosta S (2008) From crowd dynamics to crowd safety: a video-based analysis, arXiv:0810.4590
Kraus K, Harley IA (2007) Photogrammetry: geometry from images and laser scans, vol 1, 2nd edn.
Lempitsky V, Zisserman A (2010) Learning to count objects in images. Advances in neural information processing systems (NIPS), number 23, pp 1324–1332
Liu X, Song W, Zhang J (2009) Extraction and quantitative analysis of microscopic evacuation characteristics based on digital image processing. Physica A 388(13):2717–2726
Perko R, Schnabel T, Fritz G, Almer A, Paletta L (2013) Airborne based high performance crowd monitoring for security applications, scandinavian conference on image analysis (SCIA), vol 7944, pp 664–674. Springer LNCS
Steffen B, Seyfried A (2010) Methods for measuring pedestrian density, flow, speed and direction with minimal scatter. Physica A 389(9):1902–1910
Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. J Stat Mech: Theory Exp
Zach C, Pock T, Bischof H (2007) A duality based approach for realtime TV-L1 optical flow. In: Symposium on pattern recognition (DAGM), pp 214–223