DENSITY AND MOTION ESTIMATION OF PEOPLE IN CROWDED ENVIRONMENTS BASED ON AERIAL IMAGE SEQUENCES

Stefan Hinz
Institute of Photogrammetry and Remote Sensing
Universität Karlsruhe (TH), 76 128 Karlsruhe, Germany

Commission III/5

KEY WORDS: People tracking, Aerial Image Sequences, 3K camera

ABSTRACT:
Monitoring the behavior of people in complex environments has gained much attention over the past years. Most of the current approaches rely on video cameras mounted on buildings or pylons and individuals are detected and tracked in these video streams. Our approach is intended to complement this work. We base the monitoring of people on aerial camera systems mounted on aircrafts, helicopters or airships. This imagery is characterized by a very large coverage so that the distribution of people over a large field of view can be analyzed. Yet, as the frame rate of such image sequences is usually much lower compared to video streams (only 3 up to 7Hz), tracking approaches different from optical flow or KLT-tracking need to be employed. We show that reliable information for the density of groups of people, their activity as well as their locomotion can be derived from these kind of data.

1. INTRODUCTION

Monitoring the behavior of people in crowded scenes and in complex environments has gained much attention over the past years. Most of the current approaches rely on video cameras mounted on buildings or pylons and individuals are detected and tracked in these video streams. Pioneering work on tracking human individuals in terrestrial image sequences can be found, e.g., in (Rohr, 1994; Moeslund & Granum, 2001). While this work focuses on motion capture of an isolated human, first attempts to analyze more crowded scenes are described in (Rosales & Scarloff, 1999; McKenna et al. 2000). Such relatively early tracking systems have been extended by approaches integrating the interaction of 3D geometry, 3D trajectories or even intentional behavior between individuals (Zhao & Nevatia, 2004; Yu & Wu, 2004; Nillius et al., 2006; Zhao et al., 2008). Advanced approaches – based on so-called sensor networks – are able to hand-over tracked objects to adjacent cameras in case they leave the current field of view so that a quite comprehensive analysis on the monitored scene is possible. The work of (Kang et al., 2003) exemplifies this kind of approaches. Instead of networks of cameras, moving platforms like unmanned airborne vehicles can also be used, as e.g., shown in (Davis et al., 2000).

Nonetheless, the main bottleneck of these approaches is the limited coverage in case of monitoring big events in a large venue such as rock concerts, public viewing events (as e.g. during soccer world cup), and big folk festivals, which may cover several square-kilometer.

Our approach is intended to complement the above work. We base the monitoring of people on aerial camera systems mounted on aircrafts, helicopters or airships. This imagery is characterized by a very large coverage so that the distribution of people over a large field of view can be analyzed. Yet, as the frame rate of such image sequences is usually much lower compared to video streams (several Hz), tracking approaches different from typical optical flow or KLT-tracking need to be employed.

2. SENSOR AND DATA CHARACTERISTICS

For developing and testing our approach, we used aerial image sequences provided by DLR's 3K multi-head camera system (Kurz et al., 2007). This system consists of three non-metric off-the-shelf cameras (the current version of 3K is equipped with three digital Canon EOS 1Ds Mark II, 16 MPixel each, see Fig. 1). The three frame cameras can be aligned in across-track or along-track with one camera pointing in nadir direction and two in oblique direction. For achieving large ground coverage, the cameras are usually mounted in across-track, which leads to a maximum field of view of approx. 110°. Hence, within two minutes an area of roughly 10 km x 8 km can be covered. Ground sampling distance and effective swath width are depending on the flight altitude and typically range between 15cm – 50cm and 2.5km – 8km, respectively. Prerequisite is a high accuracy of the orthorectification process, which requires a self-calibration of multi-head camera system (Kurz et al., 2007).

The system can be operated in different mapping or traffic monitoring modes. Depending on the set-up, the acquisition of high resolution images, colour and wide-area monitoring is feasible, even at low flight altitudes below clouds. To provide the basis for near-realtime mapping, the system is coupled with a realtime GPS/IMU navigation system, which enables accurate direct georeferencing. Image sequences can be taken with a frame rate of 3Hz in continuous mode and even higher in burst mode (3-4 frames with framerate up to 7Hz, followed by a one-second gap for read-out). This allows also the monitoring of moving objects such as vehicles, ships or even humans (see e.g. (Hinz et al. 2007; Zeller et al., 2009; Kurz et al., 2009)).

Figure 2 depicts a single shot taken by the nadir head of the 3K system at a flight altitude of 2000m, as it has been used in the following experiments. The image covers almost the complete Oktoberfest area at "Theresienwiese" in Munich, Germany. Two bursts of the entrance area are shown in Figure 3. Each burst consists of 3 images taken with a framerate of 3Hz, the left one captured in the morning and the right one captured at noon. The different densities of humans are clearly visible. In addition, Figure 4 visualizes a detailed view on two consecutive frames. As can be seen, the spatial resolution of these data is obviously rather limited. Motions of people can nonetheless be identified, especially in the left part of this cut-out.

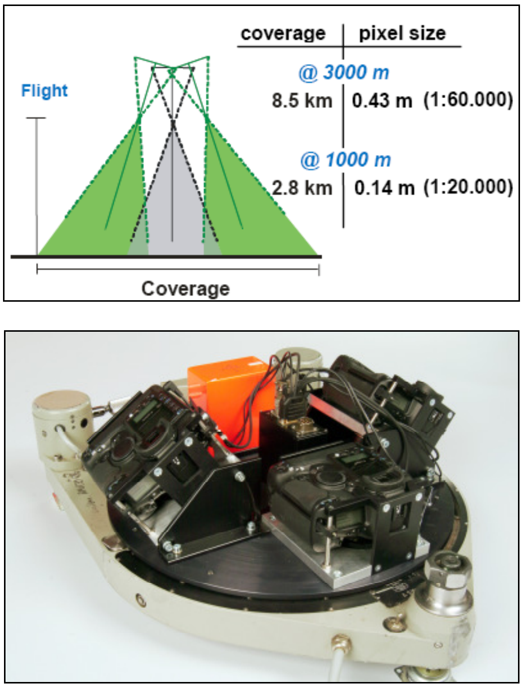

Fig. 1: Top: Image acquisition geometry; tilt of the sideward looking cameras is approx. 35°. Bottom: camera system consisting of three Canon EOS 1Ds Mark II, integrated in a ZEISS aerial camera mount including an IGI gyro System (orange) for attitude measurements.

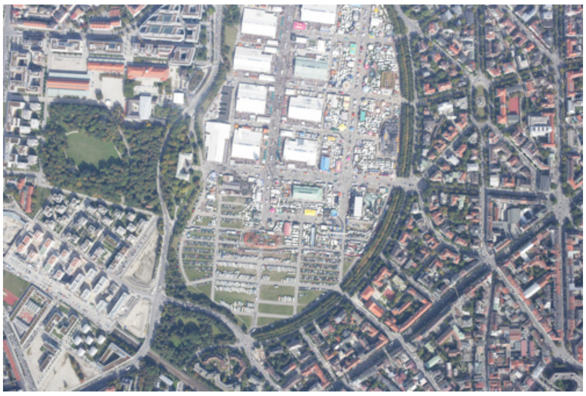

Fig. 2: Color image of the Oktoberfest area "Theresienwiese" in Munich, Germany, taken by the nadir head.

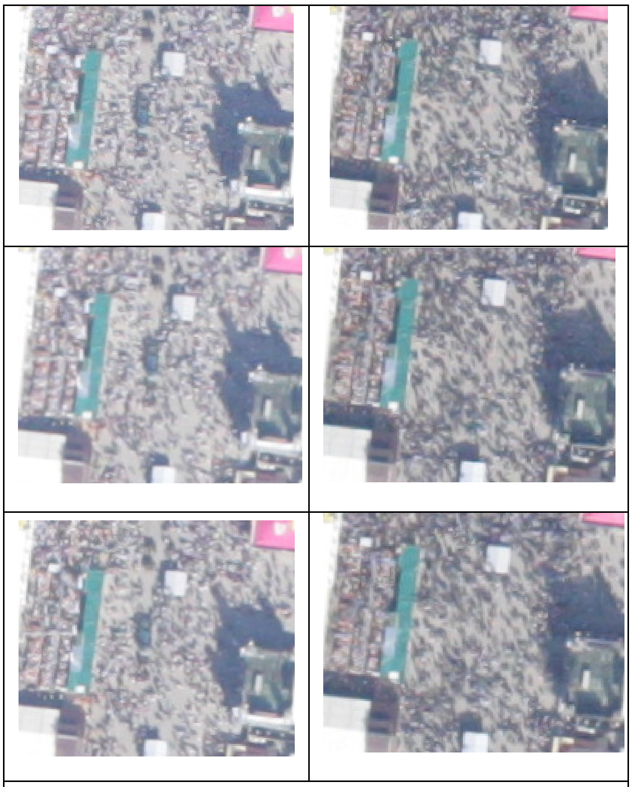

Fig. 3: Two image sequences taken with 3K's burst mode and framerate of 3Hz. Left column: scene at morning; right column: same scene at noon.

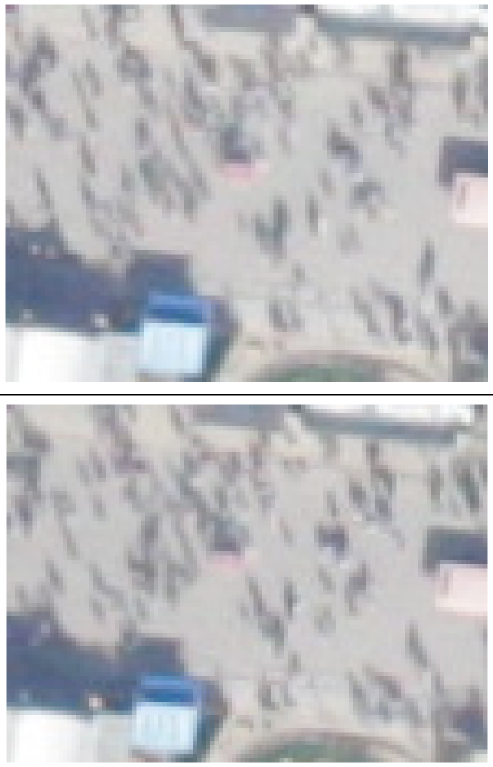

Fig. 4: Two consecutive frames of a sub-scene indicating the relatively coarse resolution, but also the capability of identifying motions of people (see esp. left part of cut-out).

3. METHOD

3.1 Overview

The estimation of image parameters that shall indicate the local density of groups of people as well as their locomotion is based on analysing spatial and temporal patterns in an image sequence or a burst sequence. In the following, the term “image sequence” relates to the complete sequence comprising various bursts and also various passes of the aircraft, whereas “burst sequence” means 3-4 consecutive frames taken during approximately 1sec.

Pre-processing comprises the determination of camera parameters and image orientations as well as co-registration of the whole image sequence and overlaying Regions-of-Interest, e.g., taken from a geo-database. These steps are assumed to be done beforehand. Then, the background of each image is estimated by analysing the gray value histogram of the burst sequence a particular image belongs to (Sect. 3.2). For the complementing region, image texture parameters are determined, which indicate the local people density (Sect. 3.3). This analysis is extended into space-time domain in Sect. 3.4, to determine the temporal agitation (i.e. people activity) present in the foreground. Finally, their motion is determined by matching image patches over images pairs of a burst sequence (Sect. 3.5).

It is important to note, that all parameters estimated with the algorithms described in Sect. 3 relate purely to image objects. Their transition into real-world object parameters shall be established by samples taken from terrestrial cameras, which calibrate the functional relationship between image parameters and object parameters. The concept of this kind of model calibration is described in the outlook in Sect. 4.

3.2 Spatio-temporal Background Estimation

Once the image sequence has been co-registered, a common approach to estimate background pixels for each image of an images sequence is the pixelwise analysis of gray value features in time domain. A simple and often used approach is to calculate the temporal median for each pixel. Time intervals, in which a certain pixel was covered by a moving object, can thus be detected by simple differencing and thresholding, as such gray values are regarded as outlier in time domain. This approach however assumes that the background is visible in the majority of frames and the illumination conditions remain approximately constant. The condition that background should be visible in the majority of frames is hardly fulfilled in case of crowded environments. To overcome this, we conduct a coupled analysis of space and time domain and estimate background pixels by introducing an a-priori likelihood function of the expected background gray value distribution.

Let g be the gray values of an image i, M_b the number of images per burst b, $p_i(g)$ the histogram of a particular image and $p_{bg}(g)$ the assumed prior background distribution, then the dominating gray value $g_{bg,b}$ of the background of a particular burst is determined by

$$g_{bg,b} = \arg\max_g \left(p_{bg}(g) \cdot \sum_{i=1}^{M_b} p_i(g) \right)$$

whereby normalization constants are neglected and $p_{bg}(g)$ is a handcrafted windowing function describing the expected brightness distribution of asphalt and concrete. The final background region $R_{bg,i}$ for each image is then determined by a regiongrowing algorithm initialized at all pixels (r_0, c_0) fulfilling the condition

$$(r_0, c_0) = \left(g(r,c) \in \left[g_{bg,b} - g_{tol}, g_{bg,b} + g_{tol} \right] \right)$$

and continuing the growing as long as the mean gray value $\overline{g_{R_j}}$ of the Region at interation step j does not exceed a predefined tolerance g_{tol}, i.e.

$$\overline{g_{R_j}} \in \left[g_{bg,b} - g_{tol}, g_{bg,b} + g_{tol} \right]$$

The described scheme relaxes the aforementioned conditions since background pixels are assumed to dominate only the gray value distribution of the burst sequence but not the distribution of a single image. Figure 5 depicts the same cut-out as Figure 3 and shows two results of background determination – one for a simpler case and another one for a more crowded environment, in which background pixels do not dominate a single image.

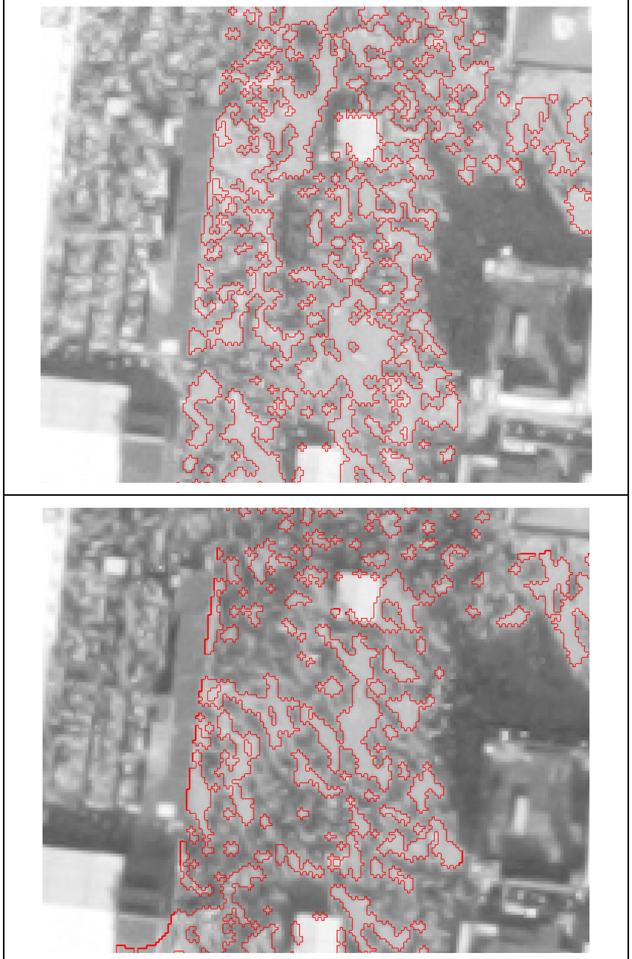

Fig. 5: Red boundary indicates background regions; less crowded situation (top), crowded environment (bottom).

3.3 People Density Estimation

The spatial density of a group of people is estimated in each image separately. A first guess about the local density can be calculated by the ratio of foreground and background pixels collected in a certain neighbourhood. However, foreground pixels usually do not entirely cover groups of people but also other features like small buildings or shadow areas. Hence, we weigh the ratio of foreground and background with the response of Laws texture filter. The result is finally smoothed to account for noise. Mathematically, this scheme can be expressed by a number of convolution operations:

$$f_{dens} = \left(\left(f_{bin} * h_{box,w}\right) \cdot \left(f_{fg} * h_{laws,ss}\right)\right) * h_\sigma$$

where f_{dens} is the resulting density image, f_{bin} is a binary image indicating foreground and background, $h_{box,w}$ is a box filter of size *w* for calculating locally the ratio of foreground and background from the binary image, f_{fg} is the original image whose domain is reduced to the foreground only, $h_{laws,ss}$ is the Laws-'ss' filter applied to foreground pixels, and h_σ is a Gaussian smoothing kernel with standard deviation σ.

Figure 6 illustrates the result of density estimation for the examples shown above. It nicely shows the increase of people from morning (left) to noon (right), especially in the left part of the cut-out. It can be also seen, however, that neighbouring objects may still influence density estimation despite of taking the texture response into account. For instance, the narrow elongated shadow area in the center of the image also produces a high texture response so that the density is overestimated in this area.

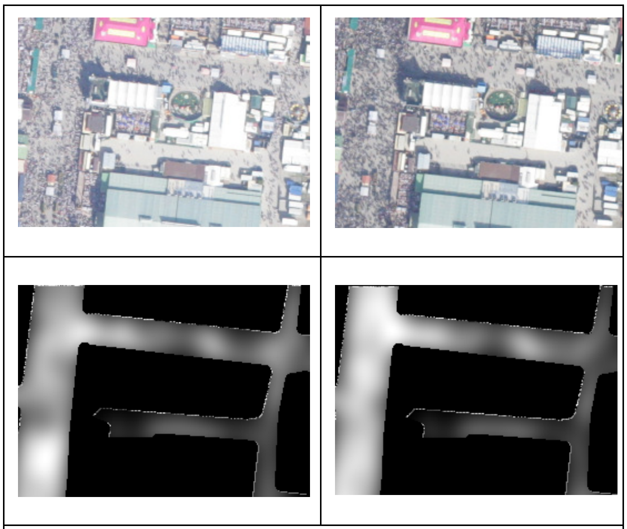

Fig. 6: Result of density estimation for morning (left) and noon (right). Light gray values indicate higher density. Especially, the higher density at noon along the left road is apparent.

3.4 People Activity Estimation

The above parameters indicate the density of groups of people but this is not necessarily linked to their activity. In fact, once the density gets higher, the ability to move freely is more and more reduced. In order to get cues about the activity of people, we analyze the temporal gray value variation within a burst sequence and make again use of texture information. As mentioned above, a simple way to detect pixels of moving objects is the calculation of the temporal gray value variance, thereby high variance indicating active regions. Though, this procedure needs a perfect co-registration of the burst sequence. Especially elevated 3D objects cause artefacts in the temporal variance response, since the underlying surface model is rarely accurate enough to completely compensate for the perspective distortion due to slightly varying viewing directions.

To reduce these effects, we weigh the temporal gray value variance with the variance of the spatial gradient directions. Regions of moving humans are typically characterized by a very inhomogeneous appearance resulting in a large variance of the spatial gradient directions. Vice-versa, man-made objects consist typically of regular structures, whose edges lead to a mono-modal or bi-modal distribution of the gradient directions. The potential influence of rectangular structures onto the gradient direction variance is thus reduced by folding all gradients into the interval $[0;\pi/2[$. Finally, the 2D mean gradient and the orientation variance are determined and spatially smoothed using a Gaussian smoothing kernel. Mathematically, this scheme can be expressed by

$$f_{act} = \mathrm{var}\left(W_{[0;\pi/2[}\left\{\nabla_{r,c} f_{\mathrm{var},t}\right\}\right) * h_\sigma$$

where f_{act} is the resulting activity map, $\nabla_{r,c} f_{\mathrm{var},t}$ the spatial gradient of the temporal variance map $f_{\mathrm{var},t}$, $W_{[0;\pi/2[}\{\cdot\}$ the wrapping operator, and *var* the variance operator.

The benefit of calculating the activity can be seen from Figure 7, in which another part of the Oktoberfest area is shown. While there is almost no activity in the upper right corner in the left example, a large group of people is walking through the narrow street in the other image. Such information could indicate a critical situation in case more people would also try to take this route – which is obviously not the case in this example.

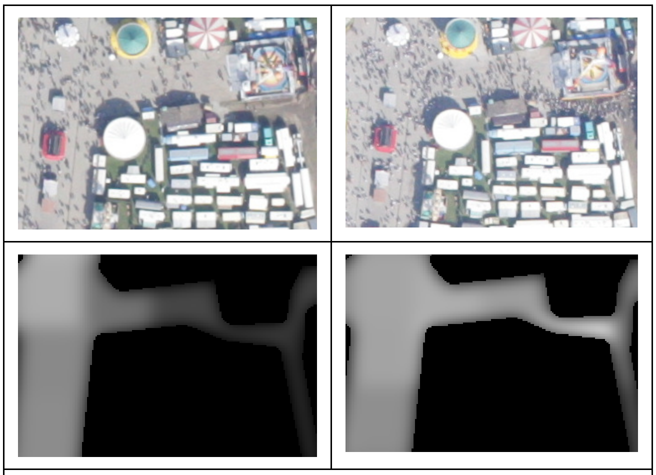

Fig. 7: Result of activity estimation based on a burst sequence. Please note in particular the junction in the upper right corner. Light gray values indicate higher activity.

3.5 Motion Estimation

In the last stage of processing, the locomotion of people is determined. As can be seen in Figure 4, tracking of individuals is hardly possible, especially in case of a crowded scene. The motion of a group, i.e. an image pattern, can be identified though.

A critical parameter for matching an image patch of a group of humans over a burst sequence is the size of the window to match. In principle, we have to expect that at least some persons move individually so that a small window containing approximately only one human would be preferable. Yet due to the relatively coarse spatial resolution, such a small window causes many multiple matches, since distinctive features describing a particular person uniquely are no more available. On the other hand, a large window may contain some persons moving in diverging directions, which would also lead to wrong or biased motion vectors. Empirical tests have shown that a window size covering approx. of 4m² on ground is a reasonable compromise, since typically only 1 – 4 humans are present in such an area. Because consecutive images are taken with time intervals of only 0.3sec, we compensate diverging motion directions of humans by a slightly coarser scale for matching, thereby accepting inaccuracies of motion determination. Future research will be directed towards developing a specific deformable model in order to match groups of humans over images.

We utilize the shape-based approach of (Steger, 2001) as matching metric. It is invariant against translations and monotonic illumination changes. Furthermore, the pyramid-based search strategy can compensate for rotation and scale changes. The similarity measure $\gamma(r,c)$ is defined as the average vector product of the gradient directions of the template and the search image

$$\gamma(r,c)=\frac{1}{n}\sum_{i=1}^{n}\frac{\left\langle d_i^m, d^s_{(r+r_i,c+c_i)}\right\rangle}{\left\|d_i^m\right\|\cdot\left\|d^s_{(r+r_i,c+c_i)}\right\|}$$

where n is the number of pixels for which the gradients have been calculated, d is the gradient direction vector in the model image m and search image s, respectively, $< \cdot >$ is the dot product and $\| \cdot \|$ is the Euclidean norm.

Figure 8 visualizes the whole area under investigation as well as two cut-outs showing details of motion determination. The windows to match are selected at regularly distributed position in the RoI. Each line represents a motion vector between two consecutive images determined at these positions. The "homogeneous" motion of people around the corner can be nicely seen in the left cut-out. The right cut-out illustrates the motion behavior in a crowded environment. Compared to the less crowded scene on the left the motion is slower, as people stand much closer to each other.

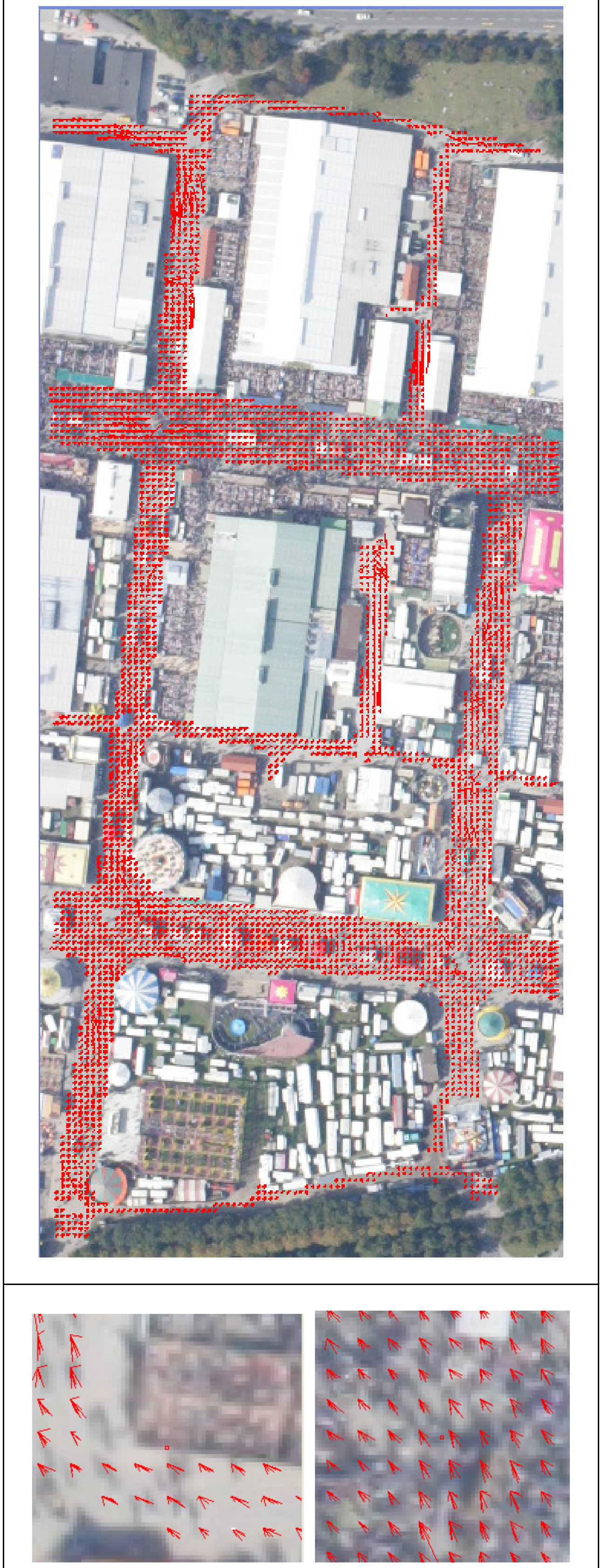

Fig. 8: People locomotion determined from consecutive image pairs. Each line represents a motion vector.

4. DISCUSSION AND OUTLOOK

We have shown that the estimation of people density and their locomotion is possible by using aerial image sequences. A careful visual analysis indicates that motion direction and velocity are determined reasonably. Similar qualitative conclusions can be drawn for people density and activity. Still missing is a quantitative evaluation. Besides this, there is much room for future research:

- Up to now, the results are image parameters that need to be transformed into object-related parameters. As single persons can hardly be determined, we believe that establishing an empirical model between image and object parameters, e.g. by numerical approximation methods like neural networks, is a more promising way than an explicit model.

- To acquire the necessary (offline) training parameters and potential (online) calibration parameters for such models, training and test data need to be collected. We plan to mount terrestrial cameras – in particular Range-Imaging (RIM) cameras – on selected position. Such cameras acquire depth and IR-reflectance images with a single shot and carry thus many advantages for 3D tracking, because the 3D information is immediately available and no stereo reconstruction needs to be done in complex environments. A limitation of these RIM cameras is their limited range of unambiguous distance measurement, since the generation of the depth image is based on the well-known continuous-wave principle. First investigations of (Jutzi, 2009) show, however, that the range ambiguity can be resolved by phase-unwrapping principles adapted from the processing of interferometric data of synthetic aperture radar.

- With the same technique, also quantitative reference data can be collected at selected sample positions.

- Further research will concentrate on the improvement of motion determination. To this end, deformable models for people tracking need to be developed and embedded into an iterative coarse-to-fine search strategy.

REFERENCES

Davis, L., Philomin, C., Duraiswami, R., 2000. Tracking Humans from a Moving Platform," Proceedings of International Conference on Pattern Recognition, vol. 4, pp. 171-178.

Hinz, S., Lenhart, D., Leitloff, J., 2007. Detection and Tracking of Vehicles in Low Frame Rate Aerial Image Sequences. International Archives of Photorgrammetry, Remote Sensing and Spatial Information Sciences, 36(1/W51), on CD.

Jutzi, B., 2009. Investigations on ambiguity unwrapping of Range Images. Submitted to ISPRS Workshop "Laserscanning'09", Paris, France.

Kang, J., Cohen, I., Medioni, G., 2003. Continuous Tracking within and across Camera Streams. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 267-272, 2003.

Kurz, F., Müller, R., Stephani, M., Reinartz, P., Schröder, M., 2007. Calibration of a wide-angle digital camera system for near real-time scenarios. International Archives of Photorgrammetry, Remote Sensing and Spatial Information Sciences, 36(1/W51), on CD.

Kurz, F., Rosenbaum, D., Thomas, U., Leitloff, J., Palubinskas, G., Zeller, K., Reinartz, P., 2009. Near real time airborne monitoring system for disaster and traffic applications. These proceedings, on CD.

McKenna, S.J., Jabri, S., Duric, Z., Rosenfeld, A., Wechsler, H., 2000. Tracking Groups of People," Computer Vision and Image Understanding 80(1): 42-56.

Moeslund, T.B., Granum, E., 2001. A Survey of Computer Vision- Based Human Motion Capture, Computer Vision and Image Understanding 81: 231-268.

Nillius, P., Sullivan, J., Carlsson, S., 2006. Multi-Target Tracking-Linking Identities Using Bayesian Network Inference," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2187-2194.

Rohr, K. 1994. Towards Model-Based Recognition of Human Movements in Image Sequences, Computer Vision, Graphics, and Image Processing: Image Understanding 59(1): 94-115.

Rosales, R., Sclaroff, S., 1999. 3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions, Proceedings of IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 117-123.

Steger, C., 2001. Similarity measures for occlusion, clutter, and illumination invariant object recognition. In: B. Radig and S. Florczyk (eds.) *Pattern Recognition,* DAGM 2001, LNCS 2191, Springer Verlag, 148–154.

Yu, T., Wu, Y., 2004. Collaborative Tracking of Multiple Targets. Proceedings of IEEE Conference Computer Vision and Pattern Recognition, vol. 1, pp. 834-841, 2004.

Zeller, K., Hinz, S., Rosenbaum, D., Leitloff, J., Reinartz, P., 2009. Traffic Monitoring without single car detection from optical airborne images. These Proceedings, on CD.

Zhao, T., Nevatia, R., 2004. Tracking Multiple Humans in Complex Situations. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(9): 1208-1221.

Zhao, T., Nevatia, R., Wu, N., 2008. Segmentation and Tracking of Multiple Humans in Crowded Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(7): pp. 1198-1211.