PFG 2013 / 2, 0083 – 0092
Stuttgart, April 2013

Article

A Stereoscopic Approach for the Association of People Tracks in Video Surveillance Systems

MORITZ MENZE, TOBIAS KLINGER, DANIEL MUHLE, Hannover, JÜRGEN METZLER, Karlsruhe & CHRISTIAN HEIPKE, Hannover

Keywords: video surveillance, stereo vision, people tracking, camera network

Summary: This article describes the application of stereoscopic analysis to typical image pairs from a surveillance camera network. An approach is presented that establishes correspondences between people detections across adjacent views and derives an estimation of body height for each person in the overlapping parts of the camera views. Dense image matching is applied to short stereoscopic sequences and the results are incorporated in a subsequent monocular tracking to improve the positioning accuracy. The method does not depend on a dedicated stereo setup of the camera network but is applicable to suitable image pairs in addition to monocular people detection and tracking. Based on realistic image sequences, the performance of the proposed approach is evaluated and compared to a current method for appearance-based data association.

Zusammenfassung: *Ein stereoskopischer Ansatz zur Zuordnung von Personenpfaden in Multi-Kamera Überwachungssystemen.* Dieser Artikel beschreibt die Anwendung eines stereoskopischen Analyseverfahrens auf typische Bildpaare eines Videoüberwachungssystems. Es wird ein Ansatz vorgestellt, der mit Hilfe der dichten Bildzuordnung aus wenigen Stereoansichten von Personen eine Größenschätzung ableitet und eine zuverlässige Übergabe von verfolgten Personen zwischen benachbarten Kameras ermöglicht. Das Verfahren ist nicht auf den Einsatz spezieller Stereosysteme angewiesen, sondern kann in Ergänzung monokularer Methoden in bestehenden Kameranetzen eingesetzt werden. Experimentelle Untersuchungen mit realistischen Bildsequenzen zeigen die Leistungsfähigkeit des vorgestellten Verfahrens verglichen mit denen weiterer Ansätze zur Übergabe von Personenpositionen zwischen verschiedenen Überwachungskameras.

1 Introduction

The automated analysis of surveillance videos is an area of active research, primarily within the Computer Vision community. Given current pan-tilt-zoom (PTZ) cameras equipped with mechanical control as well as on-board computers, research focuses for example on the development of self-organising smart camera networks (BELBACHIR 2010, JÄNEN et al. 2011). Major challenges in terms of image analysis are people detection and tracking as well as the reliable association of trajectories from individual cameras across multiple views. The latter indicates a need for the consistent handling of objects in a common reference frame in order to produce suitable data for wide area analyses (COLLINS et al. 2001). The more accurate the object coordinates of people in the scene are known, the more detailed analyses can be conducted with respect to motion patterns or interactions between tracked people.

In this paper we present an approach that generates consistent global tracks of people in non-crowded scenarios. The trajectories are calculated in a common reference frame from observations of multiple surveillance cameras. An important step in our work is the estimation of body height as well as a reliable association of tracks across partly overlapping views. For that purpose, stereoscopic analysis is applied to overlapping parts of images which are either generated randomly while scanning wide areas with several PTZ cam-

www.schweizerbart.de
DOI: 10.1127/1432-8364/2013/0160
1432-8364/13/0160 $ 2.50

eras or result from an appropriate reconfiguration of the sensor network in order to focus on areas of special interest, for example as a result of saliency detection. The presented module can be employed in addition to monocular analysis whenever the necessary preconditions are met. To demonstrate the benefit of the proposed approach we process a pair of partially overlapping image sequences according to the described setup and compare our results to manually measured reference data as well as an appearance-based method for people re-identification.

The remainder of the paper is structured as follows. In the next section we summarise previous work. Section 3 contains the strategy of our approach. It is arranged in subsections on basic geometric relations, monocular people detection and tracking, image matching and data association across multiple views. Section 4 shows experimental results derived from a realistic video sequence. Conclusions and an outlook on further work are given in section 5.

2 Related Work

In this section the state-of-the-art in data association within and across camera views is briefly reviewed. It references related work on stereo vision in surveillance applications and discusses approaches to pose estimation in the same domain.

Data association in the context of video surveillance aims at the concatenation of corresponding observations between temporally or spatially adjacent video frames. Features of tracked people that are evaluated to establish correspondences can roughly be categorised into either appearance-based or spatiotemporal features. Appearance-based features like colour histograms, descriptors of interest points or combinations of both are discussed in Doretto et al. (2011). Especially for wide baselines the appearance of one and the same object may vary significantly in the views of adjacent cameras, so that it is error-prone to use only appearance-based features for the association. Another approach to establishing correspondences is the comparison of spatiotemporal features like position and velocity (Orwell et al. 1999). Since either of the methods suffers from individual shortcomings, combined approaches were introduced. Cai & Aggarwal (1999) hand over targets between adjacent cameras predicting the position of the target in an adjacent view for re-identification. Javed et al. (2008) model probabilities of people walking certain paths and fuse this geometrical information with appearance cues using a maximum likelihood framework.

Due to the challenging task of consistent tracking, several authors gather additional depth information about the observed scene by applying stereoscopic analysis. Because stereoscopic image matching is often regarded as the central component in surveillance systems, the sensor networks are designed to fulfil the requirements of stereo approaches. The pairwise installation of PTZ cameras (Zhou et al. 2010) provides image pairs with short baselines. Dedicated stereo devices, as used in Darrell et al. (2000), Haritaoglu et al. (1998) and many other publications, capture synchronised image pairs that are processed on specialised hardware. Although the advantages of high-frequency depth maps for people detection and tracking is shown (Schindler et al. 2010), a dedicated system design leads to additional costs that, from our point of view, are not necessary, when applying stereoscopic analysis to camera networks.

In contrast, we propose the usage of stereo vision in PTZ sensor networks where overlapping fields of view are not predefined, but selectively available during short periods of handover, i. e. data association between adjacent cameras. This makes the approach applicable to existing systems of spatially distributed cameras. Nevertheless, there is still the additional need for monocular tracking where the stereoscopic approach cannot be applied.

In camera networks conceptually two approaches exist to estimate the object position in a common coordinate reference frame. The monocular position estimation can produce planar coordinates given the ground plane and the orientation parameters, either by directly measuring the intersection of objects with that plane (Collins et al. 2001) or by incorporating assumptions about the object height (Zhao & Nevatia 2004). Directly observing the point of intersection is often not possible due to occlusions, thus the observation of head tops and the incorporation of default object height is a widely used approximate solution. A second, more robust approach to position estimation is the use of multiple cameras. Eshel & Moses (2010) estimate body height calculating correspondences between multiple homographies at discrete height levels. Methods of the second category rely on the simultaneous observation of the objects throughout the entire scene, which is often not feasible due to the high amount of required resources, at least in large camera networks. Since we extract height from short stereo sequences during handover and subsequently apply monocular tracking, this limitation does not hold for the presented approach.

3 Approach

3.1 *Overview*

Given a reconfigurable network of sparsely distributed PTZ cameras, we apply monocular object tracking most of the time and exploit stereo vision whenever people pass an area viewed by two or more cameras to derive a 3D point cloud of the visible surface of each person. Based on the centroid of each dense point cloud we establish reliable correspondences between individual tracks across the views. In addition, we estimate body height which is used in all subsequent frames to improve the positioning accuracy of trajectories from monocular tracking.

3.2 *Geometric Relations*

The observations of camera networks have to be transformed to a common reference frame as a prerequisite to the automated interpretation of the whole scene in object space. In video surveillance scenarios there often exists one predominant ground plane. If this is the case, an intuitive definition of a reference frame is to align its X/Y plane to the predominant ground plane in the scene. The axes are rotated so that the Z axis directly represents object height. Image coordinates (x, y) can be projected onto a plane in object space by applying the collinearity equations (1), if the height Z of the corresponding object is known or if it lies in the ground plane (Z=0).

$$X=X_0+(Z-Z_0)\frac{r_{11}(x'-x'_0)+r_{12}(y'-y'_0)+r_{13}c'}{r_{31}(x'-x'_0)+r_{32}(y'-y'_0)+r_{33}c'}$$

$$Y=Y_0+(Z-Z_0)\frac{r_{21}(x'-x'_0)+r_{22}(y'-y'_0)+r_{23}c'}{r_{31}(x'-x'_0)+r_{32}(y'-y'_0)+r_{33}c'} \qquad (1)$$

The perspective centre of the camera (X_0, Y_0, Z_0) in the reference coordinate frame and the elements r_{ij} of the rotation between image and reference frame are determined via spatial resection from ground control points within the scene. The image coordinates of the principal point $x‘_0$, $y‘_0$ and the principal distance c‘ are assumed to be known from camera calibration. Zoom functionality is not taken into account in this paper because we aim at covering wide areas given a sparse network of cameras. Therefore, images are captured at lowest zoom level.

The influence of an erroneous object height on the projected position in object space is illustrated in Fig. 1. In common surveillance scenarios incorrect Z values result in a horizontal bias in the viewing direction of the camera. The partial derivatives of the inverse collinearity equations describe the impact of errors in the parameters. The most important factor is the body height, i. e. Z in (1).

Given a maximum deviation of 0.2 m from the default height (we use 1.72 m, see below) the maximum bias in imaging direction varies from 0.5 m to 1.1 m depending on our setup and imaging distances.

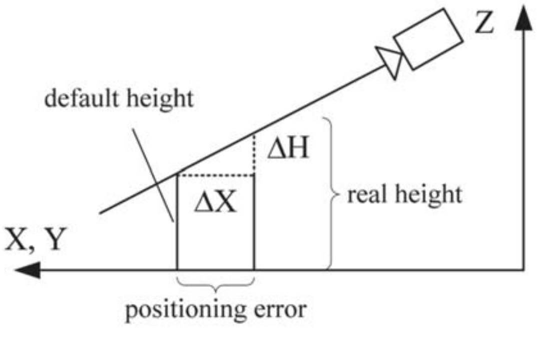

Fig. 1: Influence of height on target localisation from a monocular view.

3.3 *Detection and Tracking in Monocular Views*

People detection aims at distinguishing people from the image background. Tracking describes the process of establishing temporal correspondences between temporally corresponding people in consecutive frames. In this work we do not concentrate on finding the optimal method for monocular tracking, but apply a combination of state-of-the-art methods that works well enough in non-crowded scenarios.

For detection we follow Dalal & Triggs (2005) and classify histograms of oriented gradients within a sliding window in either people or non-people using a support vector machine (SVM). As shown in Dollar et al. (2011), false positives in sliding window approaches remain frequent. Therefore, we validate the detections with clues from background subtraction using an improved Mixture of Gaussians approach with shadow detection (Kaewtrakulpong & Bowden 2001), and select only the detections that have a sufficiently large overlap with any foreground region. Assuming a static background, valid for most surveillance scenarios, we successfully avoid false positive detections with this strategy. However, misplaced detections occur if background structure causes positive classification of a detection window which is also labelled as foreground (Fig. 2). Such incorrect detections are eliminated during data association.

We obtain temporal trajectories by finding the closest match of features between detections in adjacent frames. A greedy approach is applied, combining spatiotemporal and appearance-based features: For each target an appearance model and a motion model are set up. The appearance of each detection is described in terms of the hue histogram of the corresponding foreground blob. An appropriate similarity measure is given by the Bhattacharyya distance between histograms (Bhattacharyya 1946). After successful association the appearance model is updated by incorporating the appearance of the associated detection, which makes the tracker robust against small appearance changes. The detection with the highest similarity is associated to a trajectory, if it also fits the motion model, i. e. considering spatial position and gating the search space.

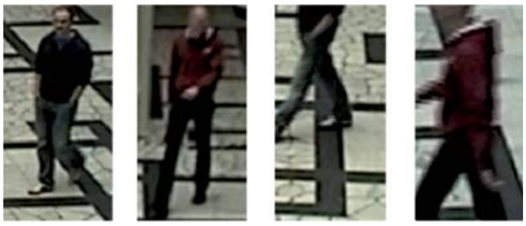

Fig. 2: People detections, resized to unit height. Two correct detections (left) and two incorrect ones (right).

A spatiotemporal description of each tracked object is realised by a Kalman Filter based on the planar position and velocity on the ground plane. For estimating the position from a monocular view one often cannot rely on the visibility of a clearly defined intersection point of the person with the ground due to occlusions, shadows and the fact, that while walking the feet obviously do not always touch the ground. The upper point of a standing person, in contrast, is likely to be identifiable in such situations. Since in monocular tracking the person‘s height is not inherently known, a default height is assumed for (1). For our experiments we have chosen a height of 1.72 m which is the average body height of a German adult (Statistisches Bundesamt 2009). As shown in section 3.1 height influences target localisation in the imaging direction. This does not necessarily degrade the performance of monocular tracking, since consecutive observations are influenced in the same systematic way. On the contrary, association across views is handicapped because the viewing directions of different cameras are generally not aligned.

The described approach can be transferred to the problem of data association across overlapping views. Only the gating of the search space has to be adapted from a temporal motion model to simultaneous observations.

3.4 *Dense Stereo Matching and Feature Extraction*

Another approach to data association across multiple views is based on dense image matching. The general idea is to deduce the geometric structure of the scene from a pixel- or region-based comparison of radiometric features in simultaneously captured images. The resulting disparity map then contains a pixelwise data association between the images.

Since we are primarily interested in pedestrians moving in the scene, the input to the matching algorithm is reduced to the foreground regions of the corresponding images inside the person tracker's bounding boxes, available from the people detection step. This intrinsically eliminates many potential mismatches. For each of the foreground regions dense stereo matching generates a disparity map that can be re-projected to object space as a point cloud on the visible surface of the respective person. As an example Fig. 3 shows two foreground patches from the input data next to the resulting disparity map and the derived point cloud, textured with colour information from the left image.

A prerequisite for successful image matching of moving objects are temporally aligned image pairs. In the absence of externally triggered cameras, we select images based on their timestamp, dropping frames without a corresponding partner. Camera clocks are synchronised using the same Network Time Protocol (NTP) server.

Corresponding images are normalised to epipolar geometry to reduce the search space in the matching procedure. The necessary image orientation is derived in a separate, automatic calibration step.

Consistent disparity maps for noisy and low textured images, as frequently encountered in surveillance footage, can be obtained by introducing smoothness constraints in the matching algorithm. An efficient implementation of such an optimisation is employed in SGM (Semiglobal Matching, Hirschmüller 2008) where a global cost function combines local matching costs with two penalty terms for different changes in disparity. Overall matching costs are minimised by means of dynamic programming along several paths performed for each pixel. The matching procedure which uses the Open Source Computer Vision library (OpenCV 2012) is based on SGM: the optimisation of an energy function integrates the described smoothness constraints and pixel-wise local matching costs, the latter being calculated as the sum of absolute differences (SAD) in a 3 × 3 window. This method is chosen due to its higher computational efficiency compared to dedicated wide baseline approaches like the DAISY descriptor (Tola et al. 2010) that allows for larger angles of convergence.

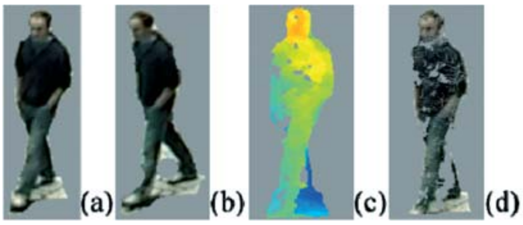

Fig. 3: Input regions for dense image matching (a, b), colour coded disparity map (c) and coloured 3D point cloud (d).

Each disparity is validated by a backmatching consistency check as proposed e.g. by Hannah (1989). Pixels that are matched from left to right have to be confirmed by a corresponding match from right to left.

After calculating the disparity image for each foreground region, pixels inside the bounding boxes of the detector are re-projected to 3D object space and transformed to the common reference frame. From the resulting point clouds, which in our setup consist of 1500 up to 3500 points on the visible surface of each person, the object height is extracted as the mean Z value of the five topmost points inside an axis-aligned square buffer of 0.8 m side length around the median X and Y coordinates corresponding to a specific bounding box.

3.5 *Data Association across Multiple Views*

Whenever a stereo image pair is processed, matching results are employed to establish correspondences between trajectories in the respective views. Globally tracked objects are mostly instantiated from a single view and associated to corresponding detections in adjacent cameras when the object enters the overlapping regions of the views.

To establish correspondences, the centroids of the three-dimensional point clouds from the matching step are re-projected to the images. Note that each point cloud corresponds to a detected person. In the image domain it is tested whether the projected point lies inside one of the bounding boxes of the people detector in the corresponding frame of the stereo image pair. If this is the case, the link is stored in a global data-structure representing the tracked object.

Correspondences are only established for unambiguous associations; the results are discarded if the projected centroid falls into multiple boxes in an image due to overlapping detections. Single misses due to such occlusions are tolerable if the overlapping region of the images is large enough to produce several synchronous image pairs showing the person. Each of the pairs is processed individually and is analysed with respect to possible associations. The presented approach conceptually works with only one valid association. However, multiple associations of the same tracks increase reliability.

This procedure relies on simultaneous detections in both views and does not directly support detection by providing clues to the presence of people in the scene. Such integrated procedures are used in multiple view detection approaches, e.g. ZHAO et al. (2005), whereas this paper focuses on handover in short sub-sequences.

4 Experiments

4.1 *Overview*

In this section, the proposed approach is evaluated on realistic surveillance footage collected in an indoor setting by our prototype system. First, the dataset is described. The remaining subsections evaluate the performance of the proposed approach with respect to data association and the incorporation of the body height estimation.

4.2 *Dataset*

Although there are several publicly available test datasets addressing people detection and tracking, e. g. performance evaluation of tracking and surveillance (PETS 2012), context aware vision using image-based active recognition (CAVIAR 2003), and video and image retrieval and analysis tool (VIRAT 2011), none of them provides suitable input to the presented approach. This is due to the fact that, although a few overlapping views exist in some of the datasets, convergence angles and scale differences exceed reasonable limits for stereoscopic analysis as described in this paper.

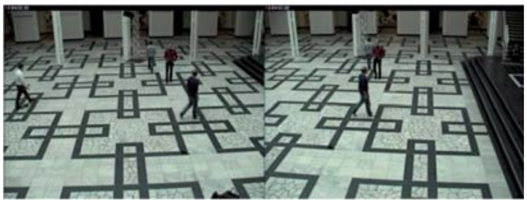

Fig. 4: Image pair from the test sequence.

Therefore, we have decided to acquire additional test data ourselves. To demonstrate handover and tracking, image sequences from two cameras are processed. Fig. 4 shows a synchronous pair of images from the experimental setup. The cameras are mounted approximately 5.7 m above the ground plane with a stereo base of 4.2 m. People pass at distances from 8.5 m to 18.5 m in front of the camera, yielding base-to-distance ratios of 0.49 to 0.23. The image resolution of each camera is 768 × 576 pixels and images are captured at a frame rate of 15 fps.

4.3 *Data Association*

Handover is evaluated on a test sequence of 1000 image pairs with 20 people passing the overlapping region. The density of people is relatively low and does not exceed 3 people crossing the region of overlap simultaneously. To be able to process all possible associations, manually labelled detections are used for the experiments in this subsection.

One computationally efficient geometric approach to data association in object space is to establish links based on the distance between detections from different views and a threshold. To compare this strategy (called "mono" in the remainder) to the stereoscopic approach, a simple metric is defined by the ratio of successfully associated vs. all possible

Tab. 1: Absolute number and percentage of correct associations on a per frame basis.

	no. correct	% correct
mono	377	74.8
stereo	466	92.5
mono+appearance	503	99.8

Tab. 2: Absolute number and percentage of correct associations on a per track basis using manually labelled detections.

	no. correct	% correct
mono	18	90
stereo	18	90
appearance	19	95
mono+appearance	20	100

associations. Given manually labelled reference data we can directly compare the performance of the monocular and the stereoscopic approach on a per frame basis. A threshold of 0.3 m was used for the mono cases. Tab. 1 shows that the stereoscopic approach clearly outperforms the simpler mono approach. However, the combination of monocular geometric positioning and appearance-based features, as described in section 3.3, yields almost flawless results for the given dataset.

To evaluate the performance of the presented approach with respect to the state-of-the-art, we processed the described test dataset with a recent appearance-based approach to people re-identification (Metzler 2012) which is based on the mean of covariance descriptors calculated from several images and does not incorporate position information. Since the appearance-based approach builds a descriptor from multiple images of each person, it cannot be compared on a per frame basis. Tab. 2 gives the results in terms of the number and percentage of correctly associated tracks.

The stereoscopic approach works almost as well as the appearance-based method which has the advantage of relatively few candidate matches in the second image. The combined spatiotemporal and appearance-based method (section 3.3) is able to associate all 20 tracks correctly. However, the results differ by one or two detected tracks only, when manually detected bounding boxes are used. Thus, it can be concluded that for this experiment, all investigated methods deliver nearly the same quality of results.

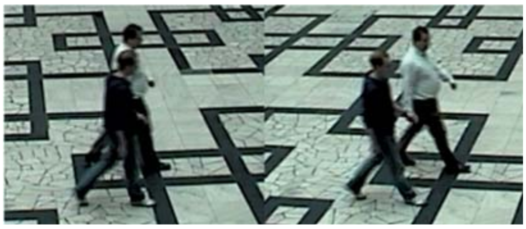

Fig. 5: Heavy occlusion hindering people detection and reconstruction.

Using the results of our detector/tracker instead of manual labelling, the overall performance of the system degrades. Due to partial or complete misses of the detector, the absolute number of possible associations becomes smaller and the reliability of data association decreases as well. For our test dataset only 17 out of 20 people are detected.

Although the stereoscopic approach can deal with a certain amount of occlusion, the reconstruction fails, just as the detector does, if people walk too close together. Fig. 5 depicts a situation where only the person in front is detected and reconstructed. Since the problems depicted in Fig. 5 stem from the detection deficiencies, similar problems occur with the other three association methods. In summary, all four association methods delivered nearly identical results for these detections.

4.4 *Incorporation of Body Height*

The effect of body height estimation is discussed based on the results of our tracker for a single trajectory. Fig. 6 shows an exemplary tracking result in the reference frame of our test site. The upper right of the figure corresponds to the upper right of the images in Fig. 4. The moving direction of the target that is tracked was from the upper right to the lower left corner in Fig. 6. The rectangle in the middle indicates the area in which the target was observed stereoscopically. Manually labelled ground truth is depicted as a blue

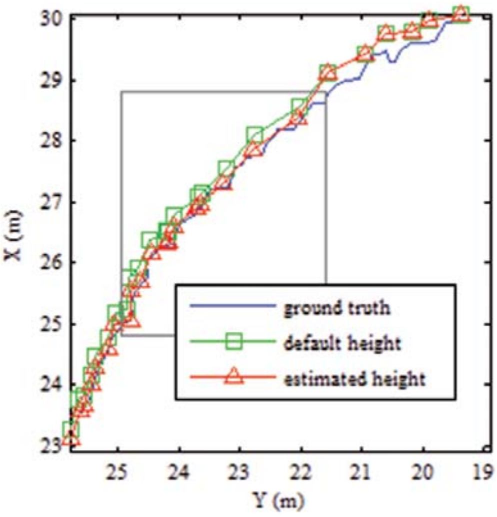

Fig. 6: Exemplary trajectory in a common reference frame. Parts inside the rectangle are processed stereoscopically.

line. The green line is the trajectory after associating the monocular tracking results from the left and right view using the described stereoscopic approach. Detections are projected to the ground plane using the average height. The red line is the trajectory updated with the height estimation from stereo processing as soon as available. None of the trajectories are filtered in the time domain. Initially, both the green and the red trajectory show a systematic offset from what is labelled as ground truth because of the default height they assume. Note that the red trajectory converges towards the reference data from the moment the target reaches the overlapping area, while the green trajectory retains its systematic offset in X direction (imaging direction).

Tab. 3 gives numerical results of a comparison between our measurements and manually labelled reference data, computed as the root-mean-squared difference between corresponding points in the ground plane. The left part gives overall results of the trajectories in Fig. 6 and therefore includes biased results from default height. The mean difference between tracking results and reference data is reduced significantly from 26 cm to 11 cm in X direction (which is the imaging direction) while differences in Y direction are not influenced. This is the expected improvement, since an error in height directly results in a biased position in imaging direction.

The overall improvement of the target localisation with respect to the reference data demonstrates the benefit of our approach. In the right part of Tab. 3, only those trajectory points with an estimated body height are compared to the respective biased results. While the mean difference in Y direction slightly changes due to the reduced number of samples, the differences in X direction are reduced by an order of one magnitude.

5 Conclusions and Outlook

We presented an approach to the generation of globally consistent trajectories from surveillance videos. Our contribution deals with the handover of tracked objects between different cameras with occasionally overlapping fields of view. By exploiting stereo vision during handover a reliable estimation of body height can be obtained. The major benefit of the approach is the increased geometric positioning accuracy during stereoscopic and subsequent monocular tracking. An improved geometric accuracy of the trajectories enables a more precise description of movements in the scene and a more detailed analysis of interactions.

The approach has been successfully tested on realistic image sequences. Positioning accuracy is improved and data association performs on a level comparable to state-of-the-art methods. The integration of a more sophisticated association procedure could further improve these results. Ambiguous associations of the stereoscopic approach due to overlap-

Tab. 3: Mean difference of tracking results and reference data.

	overall		overlapping	
	ΔX (m)	ΔY (m)	ΔX (m)	ΔY (m)
default height	0.26	0.02	0.22	0.06
estimated height	0.11	0.02	0.02	0.06

ping bounding boxes in the second image could be resolved by incorporating additional appearance-based features in such cases.

Focussing single people of interest at higher resolution will yield more detailed point clouds. Those could be analysed with respect to full body motion and action recognition. More extensive tests will be conducted in future works by integrating the approach into a wider network of self-organising smart cameras.

Our method is designed for the application in reconfigurable sensor networks with limited resources that use spatially distributed PTZ cameras to cover wide areas. In such a setup, stereo vision is only applied during the short periods of handover. The improvement in target localisation achieved by incorporating the correct body height makes our approach also suitable for systems that aim at the optimisation of global trajectories in post-processing. So far, only online applications in self-organising smart camera networks are considered that require tracking results on the fly. Applying corrections to past observations after a body height was determined may be useful for the backward tracking of people in recorded sequences.

Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (BMBF), 13N10809 – 13N10814. The support is gratefully acknowledged.

References

Belbachir, A. (ed.), 2010: Smart Cameras. – 1. edition, 404 p., Springer, New York.

Bhattacharyya, A., 1946: On a Measure of Divergence between Two Multinomial Populations. – Sankhya: The Indian Journal of Statistics **7:** 401–406.

Cai, Q. & Aggarwal, J., 1999: Tracking human motion in structured environments using a distributed-camera system. – IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (11): 1241–1247.

CAVIAR, 2003: http://homepages.inf.ed.ac.uk/rbf/CAVIAR (7.1.2013).

Collins, R, Lipton, A., Fujiyoshi, H. & Kanade, T., 2001: Algorithms for Cooperative Multisensor Surveillance. – IEEE **89:** 1456–1477.

Dalal, N. & Triggs, B., 2005: Histograms of oriented gradients for human detection. – IEEE Conference on Computer Vision and Pattern Recognition **1:** 886–893, San Diego, USA.

Darrel, T., Gordon, G., Harville, M. & Woodfill, J., 2000: Integrated Person Tracking Using Stereo, Color, and Pattern Detection. – International Journal of Computer Vision **37** (2): 175–185.

Dollar, P., Wojek, C., Schiele, B. & Perona, P., 2011: Pedestrian Detection: An Evaluation of the State of the Art. – IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (4): 743–761.

Doretto, G., Sebastian, T., Tu, P. & Rittscher, J., 2011: Appearance-based person reidentification in camera networks: problem overview and current approaches. – Journal of Ambient Intelligence and Humanized Computing **2** (2): 127–151.

Eshel, R. & Moses, Y., 2010: Tracking in a Dense Crowd Using Multiple Cameras. – International Journal of Computer Vision **88** (1): 129–143.

Hannah, M., 1989: A System for Digital Stereo Image Matching. – Photogrammetric Engineering and Remote Sensing **55** (12): 1765–1770.

Haritaoglu, I., Harwood, D. & Davis, L., 1998: W^{4}S: A Real-Time System for Detecting and Tracking People in 2 ½ D. – Lecture Notes in Computer Science **1406:** 877–892.

Hirschmüller, H., 2008: Stereo Processing by Semiglobal Matching and Mutual Information. – IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (3): 328–341.

Jänen, U., Huy, M., Grenz, C., Hoffmann, M. & Hähner, J., 2011: Distributed Three-Dimensional Camera Alignment in Highly-Dynamical Prioritized Observation Areas. – IEEE International Conference on Distributed Smart Cameras: 38–43.

Javed, O., Shafique, K., Rasheed, Z. & Shah, M., 2008: Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. – Computer Vision and Image Understanding **109** (2): 146–162.

Kaewtrakulpong, P. & Bowden, R., 2001: An improved adaptive background mixture model for real-time tracking with shadow detection. – **2nd** European Workshop on Advanced Video Based Surveillance Systems: 1–5, London, UK.

Metzler, J., 2012: Appearance-based Re-Identification of Humans in Low-Resolution Videos using Means of Covariance Descriptors. – IEEE International Conference on Advanced Video and Signal-Based Surveillance: 191–196, Beijing, China.

OpenCV, 2012: Open Source Computer Vision library, http://opencv.willowgarage.com (1.1.2013).

Orwell, J., Massey, S., Remagnino, P., Greenhill, D. & Jones, G., 1999: A Multi-agent Framework for Visual Surveillance. – International Conference on Image Analysis and Processing: 1104–1107, Venice, Italy.

PETS, 2012: http://www.pets2012.net (7.1.2013).

Schindler, K., Ess, A., Leibe, B. & Van Gool, L., 2010: Automatic detection and tracking of pedestrians from a moving stereo rig. – ISPRS Journal of Photogrammetry and Remote Sensing **65** (6): 523–537.

Statistisches Bundesamt, 2009: www.destatis.de (1.1.2013).

Tola, E., Lepetit, V. & Fua, P., 2010: DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo. – IEEE Transactions on Pattern Analysis and Machine Intelligence **32:** 815–830.

VIRAT, 2011: http://www.viratdata.org (7.1.2013).

Zhao, T. & Nevatia, R., 2004: Tracking multiple humans in complex situations. – IEEE Transactions on Pattern Analysis and Machine Intelligence **26:** 1208–1221.

Zhao, T., Aggarwal, M., Kumar, R. & Sawhney, H., 2005: Real-time wide area multi-camera stereo tracking. – IEEE Conference on Computer Vision and Pattern Recognition **1:** 976–983, San Diego, USA.

Zhou, J., Wan, D. & Wu, Y., 2010: The Chameleon-Like Vision System, – IEEE Signal Processing Magazine **27** (5): 91–101.

Addresses of the Authors:

Moritz Menze, Tobias Klinger, Dr.-Ing. Daniel Muhle, Prof. Dr.-Ing. habil. Christian Heipke, Leibniz Universität Hannover, Institut für Photogrammetrie und GeoInformation, Nienburger Str. 1, 30167 Hannover, Germany, Tel.: +49-511-762-17488, Fax: +49-511-762-2483 , e-mail: surname@ipi.uni-hannover.de

Jürgen Metzler, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung, Abteilung Videoauswertesysteme, Fraunhoferstraße 1, 76131 Karlsruhe, Germany, Tel.: +49-721-6091-453, e-mail: juergen.metzler@iosb.fraunhofer.de

Manuskript eingereicht: November 2012
Angenommen: Januar 2013