USING STEREO VISION TO SUPPORT THE AUTOMATED ANALYSIS OF SURVEILLANCE VIDEOS

Moritz Menze and Daniel Muhle

Institute of Photogrammetry and GeoInformation
Leibniz Universität Hannover
Nienburger Straße 1, 30167 Hannover, Germany
menze@ipi.uni-hannover.de, muhle@ipi.uni-hannover.de
http://www.ipi.uni-hannover.de

Commission III/1

KEY WORDS: Image Sequences, Stereoscopic Vision, Image Matching

ABSTRACT:

Video surveillance systems are no longer a collection of independent cameras, manually controlled by human operators. Instead, smart sensor networks are developed, able to fulfil certain tasks on their own and thus supporting security personnel by automated analyses. One well-known task is the derivation of people's positions on a given ground plane from monocular video footage. An improved accuracy for the ground position as well as a more detailed representation of single salient people can be expected from a stereoscopic processing of overlapping views. Related work mostly relies on dedicated stereo devices or camera pairs with a small baseline. While this set-up is helpful for the essential step of image matching, the high accuracy potential of a wide baseline and the according good intersection geometry is not utilised. In this paper we present a stereoscopic approach, working on overlapping views of standard pan-tilt-zoom cameras which can easily be generated for arbitrary points of interest by an appropriate reconfiguration of parts of a sensor network. Experiments are conducted on realistic surveillance footage to show the potential of the suggested approach and to investigate the influence of different baselines on the quality of the derived surface model. Promising estimations of people's position and height are retrieved. Although standard matching approaches show helpful results, future work will incorporate temporal dependencies available from image sequences in order to reduce computational effort and improve the derived level of detail.

1 INTRODUCTION

The increasing number of surveillance cameras and the corresponding amount of videos to be checked establish a need for automated analysis of such footage. Due to the amount of video material on the one hand and limitations in financial and technical resources on the other, human operators are not able to detect salient behaviour on-line for wider areas nor can they handle a manual reconfiguration of large sensor networks to achieve a good coverage. The completion of these tasks based on reasonable resources w.r.t the number of cameras and security personnel is one major focus of research on video surveillance systems.
Impressive progress has been made in terms of people detection and tracking, mostly relying on monocular views of the scene. Stereoscopic analysis is shown to be very helpful for detection and tracking, especially in more complex scenes or in presence of occlusions. Furthermore, dense image matching can be used to reconstruct the visible surface of people. This leads to a more robust estimation of the person's position in the scene as well as to a hint regarding it's posture enabling additional reasoning about people's actions and interactions in three-dimensional space. Up to now, stereoscopic analysis is mostly carried out using special and more expensive stereo devices or network set-ups.
Recent work on reconfigurable smart camera networks aims at self-organising structures capable of automated reconfiguration with respect to maximum coverage or focus on individual subjects. This paper investigates the employment of stereo vision algorithms using pairs of standard pan-tilt-zoom (PTZ) cameras of a reconfigurable sensor network. The presented approach and does not require special stereo devices or dedicated and possibly redundant configuration of the camera network.
The remainder of this paper is organised in four sections. It gives a brief insight into related work in section 2, presents the proposed approach in section 3, describes its experimental validation in section 4 and ends with section 5 drawing conclusions and giving an outlook to future work.

2 RELATED WORK

Using stereo vision approaches in video surveillance is a notion that regularly appears in the literature. A lot of publications make use of stereo vision devices like Konolige's Small Vision System (Konolige, 1997) which consists of mechanically aligned cameras and dedicated stereo algorithms. It is integrated into a combined detection and tracking algorithm for one static stereo camera in (Haritaoglu et al., 1998). In that publication, the disparity map is combined with an intensity based model to implement background subtraction. Removal of background is an important step in this paper as well but it will rely on colour information to extract foreground regions which are then processed by dense image matching.
(Zhao et al., 2005) employ a network of stereo cameras with overlapping fields of view which yields reliable tracking results even in complex environments but also makes excessive use of resources. Less resource-intensive is the approach presented by (Zhou et al., 2010). A pair of PTZ cameras is installed with a short baseline and utilised to generate high-resolution images of detected people as well as three-dimensional information from stereo vision. In contrast to these approaches, the presented work investigates the application of stereo vision in almost arbitrary PTZ camera networks allowing its use in existing video surveillance systems.
In this paper, we focus on the derivation of more robust and more accurate geometrical information about people classified as salient. Thus, the approach has to be embedded into a framework of components for people detection and tracking (Monari et

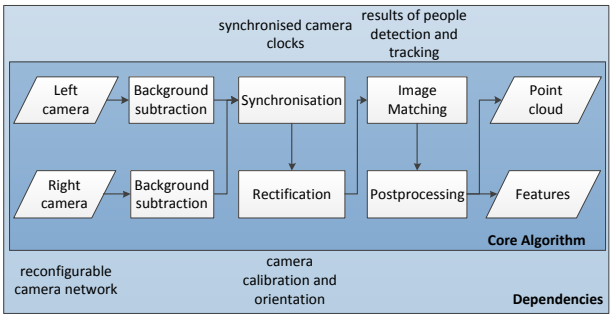

Figure 1: Core algorithm and external dependencies.

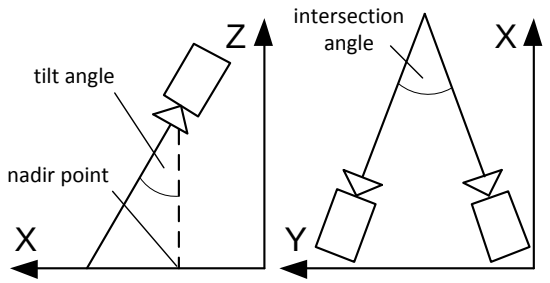

Figure 2: Imaging geometry in monocular and stereoscopic case.

al., 2009), classification of observed people into either salient or unremarkable and dynamical reconfiguration of the camera network (Jänen et al., 2011). All of these components are investigated in the joint research project CamInSens[1].

3 METHOD

Spatially and temporally aligned image pairs are the prerequisite input to standard image matching approaches. While technical details of the reconfiguration of our sensor-network are described in (Jänen et al., 2011), the major steps of our approach are presented in this section. After a comparison of monocular and stereoscopic approaches to the analysis of surveillance videos, the four major steps of our method are described, i.e. preprocessing of the individual image sequences, background subtraction, dense stereo matching and post-processing of matching results. The section ends with the estimation of measuring uncertainties. An overview of the core algorithm as well as external dependencies is given in figure 1.

3.1 Comparison of monocular and stereoscopic analyses

Given the exterior orientation of the camera, objects' positions on a known ground plane can be inferred from monocular images. Resulting positioning accuracy heavily depends on the measuring uncertainty in the image and on the tilt angle between camera and ground plane. The positioning uncertainty decreases with the imaging direction approaching the nadir point, i.e. a birds-eye view of the scene (cf. figure 2). Together with the tilt angle, geometrical positioning uncertainty decreases as well as the information content of the images since people in the scene usually cannot be identified from views of their head-tops and shoulders. Because recognition of people is a major task in the manual and automated analysis of surveillance videos, the most commonly used camera set-up is a heavily tilted one. Adapting the imaging geometry to improve recognition inevitably increases positioning uncertainty. Furthermore, tilted views of the scene can lead to partial occlusion of the targets which does not necessarily interfere with identification, since heads will remain visible, but can hinder the direct measurement of people's point of contact to the ground. Stereoscopic analysis of the scene can overcome these problems inferring people's position from a huge number of measurements on the visible surface. Its positioning accuracy mainly depends on the distance-to-base ratio and the resulting intersection angle as depicted in figure 2.

3.2 Pre-processing

Since we are using standard PTZ network cameras one of the most important preprocessing steps is to select appropriate stereo pairs from the live-streams. Given moving people, a temporal offset between the images would lead to different poses of the person in the left and right frame hindering image matching or leading to errors in depth calculation. To retrieve corresponding images, their timestamps, written to the image header on the camera when the image is actually captured, are compared and only nearest neighbours are selected as candidate pairs. The cameras' clocks are synchronised with the same NTP server to avoid offsets. Moreover, the difference between adjacent timestamps has to fall short of a threshold of one frame, i.e. $0.06\,s$ given a frame rate of $15\,fps$. As a result of this synchronisation procedure only a subset of images is actually processed which reduces the actual frame rate and leads to variable offsets between subsequent images.

For the investigations in this paper we fix the interior orientation of the cameras and apply a constant set of calibration parameters. All cameras in our network are calibrated off-line using standard methods and a calibration pattern. The exterior orientation is derived based on image features with known object coordinates. In our dataset there is a significant pattern on the floor simplifying the estimation of the exterior orientation. Since the cameras did not move while capturing images the control points' image coordinates could be measured manually. Spatial resection based on this data yields reliable results for the exterior orientation parameters.

In order to make use of parallel epipolar lines with identical y coordinates, the images are rectified to the normal case. Given this standard geometry, image matching is reduced to a one dimensional search for correspondences along the epipolar line. This set-up simplifies the selection of matching image rows and the calculation local of matching costs.

3.3 Background Subtraction

Given the application of video surveillance, the analyses as well as the input data to image matching can be reduced to people, i.e. foreground objects, in the scene. Subtracting a static background heavily reduces the amount of data to be processed and eliminates an important source of errors.

Background subtraction has been investigated intensively in recent years and elaborate methods have been proposed, giving reliable foreground regions at high frame rates. In this paper, we apply the approach described in (KaewTraKulPong and Bowden, 2001), followed by morphological closing in order to retrieve consistent foreground regions. Dense image matching is applied to the foreground regions, exemplarily shown in the two leftmost images of figure 3.

3.4 Dense Image Matching

The aim of the presented approach is to generate additional and more detailed information on people that have been classified as

[1] www.caminsens.org

Figure 3: Input to dense matching (left), colour-coded disparity map (center) and coloured point cloud (right).

salient. The basic idea is to have a closer look at these people over a limited stretch of time. To derive as many points as possible on the person's surface dense image matching is applied to the foreground regions of the rectified image pair. A consistent disparity map can be used to derive a point-wise reconstruction of the person in three dimensional object space.

To reduce computational effort and to exclude potential mismatches, image matching is performed on small parts of the images. Input to the matching algorithm are the foreground pixels inside the bounding box from a preceding people detection step. For the investigations in this paper, we make use of manually labelled bounding boxes simulating error-free output of a people detector. This region of interest in the left image is matched to the corresponding rows in the right image using a semiglobal optimisation approach (Hirschmüller, 2008). The algorithm calculates local matching costs based on the sum of absolute differences (SAD) of three-channel colour images and optimises the resulting disparity image by minimising an energy function along 16 paths ending in the current pixel. Changes of disparity in neighbouring pixels are penalised by penalty terms for small differences or larger ones respectively. Additionally, a left-right disparity check is performed eliminating implausible matches. Application of this algorithm leads to a smooth disparity map of the visible surface.

An exemplary result is shown in the center of figure 3. The disparity map is given with larger disparities coloured orange and smaller ones coloured blue. Since the cameras are mounted approximately $3\ m$ overhead and the viewing direction is slanted towards the ground plane, the disparities change slightly along the body's major axis.

3.5 Post-processing

After the calculation of the disparity map, object coordinates of the matched image points are retrieved by re-projecting valid disparity measures to the model coordinate frame of the stereo pair. Afterwards, model coordinates are transformed to the common reference frame using the left cameras' exterior orientation parameters.

The median of the resulting X- and Y-coordinates is calculated as a robust approximate value of the position belonging to the person that mainly occupies the bounding box. Partial occlusions in the image due to objects between the focussed person and one of the cameras can be handled by the presented approach based on the use of 3D information clearly separating people in object space. To exclude points on additional objects that lie within the bounding box in the image, a square buffer of $40\ cm$ side length around the median coordinates is applied to the point cloud in object space. The final position of the person is derived from the mean X- and Y-coordinates of all points inside the buffer. An estimation of the person's body height is calculated as the mean Z-value of the five topmost object points inside the buffer.

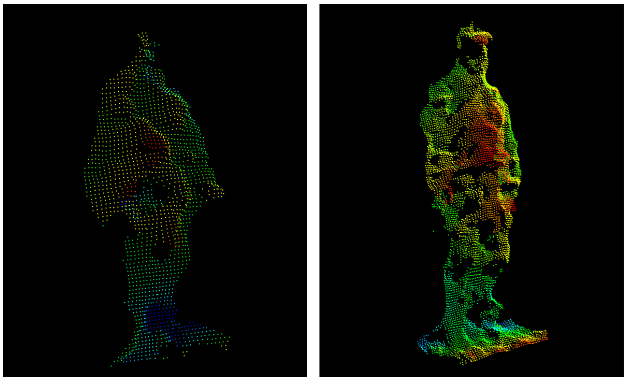

Figure 4: Exemplary point clouds of the same person in different imaging distances.

3.6 Uncertainty Estimation

To assess the derived results it is necessary to evaluate a measure of uncertainty in object space. Error propagation delivers estimations of uncertainty (cf. equation (1)) in imaging direction s_Z and perpendicular to it s_X (McGlone, 2004), neglecting the influence of the orientation parameters' uncertainties.

$$s_X = \frac{Z}{c} \cdot s_{x'},\ s_Z = \frac{Z^2}{b \cdot c} \cdot s_{px} \tag{1}$$

Z is the calculated depth of the object point and c is the principal distance, the ratio of both is the scale number. $s_{x'}$ denotes the uncertainty of image coordinates and is set to 1 pixel

Equation 1 gives uncertainties in the model coordinate frame centred in the left camera's optical centre. The variance-covariance matrix has to be transformed to the common reference frame with the same rotation as applied to the coordinates.

The limits of feasible imaging configurations are determined considering the resulting uncertainties. In this case, the relative uncertainty, i.e. the ratio between uncertainty and calculated distance, is an appropriate measure of quality. Given a fixed stereo baseline, a maximum distance can be found yielding a relative accuracy below a threshold of 0.001. For the dataset discussed in section 4 this yields a maximum distance of $13.5\ m$ between object and cameras. This consideration can be reversed to select appropriate cameras to observe a certain target.

4 RESULTS AND DISCUSSION

To evaluate the proposed approach realistic surveillance video footage is processed. PTZ cameras were installed in an atrium at the Leibniz University Hannover where typical scenes have been recorded. An exemplary stereo pair is shown in Figure 5 together with the axes of the common reference coordinate frame. Note that the origin is translated to the visible image region for visualisation. In the figure, two raw images are simply stacked, neither corrections of lens distortions nor rectification are applied. Both cameras are mounted approximately $4.7\ m$ above ground level with a stereo baseline of $4.2\ m$. An area of $10 \times 10\ m$ is covered by overlapping fields of view at a minimal distance of $6\ m$ from the cameras yielding distance-to-base ratios from 1.4 to 3.8. Our algorithm runs at an average frame rate of $4\ fps$ on the given dataset. The described methods are employed to extract people's position and height, results are compared against manually measured reference data.

Figure 5: Stereo image pair from test sequence. The coordinate system is translated into the image to show the directions of the axes. In the left image a bounding box is depicted as well as the position derived from our approach (green circle).

In our approach, people's position on the ground plane is calculated as the mean X- and Y-coordinate of the reconstructed point cloud. A monoscopic approach is to re-project the center of the bounding box's lower boundary on the ground plane. For this purpose we use manually labelled bounding boxes. Figure 6 shows the results for the person marked in figure 5 which approaches the cameras on a straight line. The upper graph shows the person's position on the ground plane, derived from the point cloud, as blue circles and reference data as red crosses. The X-coordinate corresponds to the distance between object and camera. There is an offset between both trajectories resulting from the shape of the bounding boxes. Since they enclose the limbs of the person, the monoscopic projection is biased. In the left image of figure 5 our result is projected to the image plane as a green circle and shown together with the bounding box. Note the offset between the results due to the person's foot.

The lower graph of figure 6 gives the height estimates and their individual uncertainties (blue bars) plotted against the true body height (red line). With decreasing distance to the cameras, measuring uncertainty decreases as well as the deviation from the true height. One can clearly identify the periodical movement of the head indicating a high relative measurement accuracy. (Hirasaki et al., 1999) show that the vertical translation of the head varies between $1\ cm$ and $3\ cm$ depending on walking velocity which corresponds to our results.

	Person 1	Person 2	Person 3
mean estimated height [m]	1.80	1.82	1.93
standard deviation [m]	0.02	0.04	0.03
true body height [m]	1.83	1.83	1.93

Table 1: Mean estimated body height and ground truth for different people in the scene.

To show the accuracy potential of the approach given the described dataset the mean body height derived from our measurements is compared to ground truth for three people in table 1. The standard deviation of the height estimation, calculated over all epochs, varies between 0.02 and $0.04\ cm$, the biggest deviation results for person 1. It's trajectory, as shown in figure 6, partially exceeds the maximum measuring distance of $13.5\ m$ discussed in section 3. Note that the offset between height estimation and ground truth increases significantly on the other side of this threshold. For Person 3 the results are identical to ground truth, the person crosses the scene perpendicular to the viewing direction at a relatively short distance of $8.5\ m$.

In absence of reference measurements the quality of the resulting surface models is evaluated by visual inspection. Figure 4 gives point clouds of the same person in maximum and minimum imaging distance. The total number of points is 1890 and 5936 respectively. While the left point cloud still consists of enough points to retrieve a robust estimation of the person's position the determination of the body height suffers from the heavily reduced number of points on the head.

During our investigations another dataset of the same scene was processed to assess the influence of even wider stereo baselines. Due to a different camera set-up and a reduced scale and resolution in object space the resulting point clouds are significantly smaller. Only 330 points are reconstructed on a person at an imaging distance of $20\ m$. Thus a reliable height estimation is no longer feasible but robust positioning results are retrieved for a baseline of $11.6\ m$. This shows the principal applicability of the approach to a wide range of camera set-ups.

5 CONCLUSIONS AND FUTURE WORK

The application of stereo vision to surveillance videos produces more robust results compared to monocular analysis. Estimates of people's positions in the scene and of their body height can be derived from the point cloud reconstructed on the visible surface. Both measures can be obtained from low resolution imagery. However, processing of higher resolution material would enable a more detailed representation of people in terms of an articulated body model as well as a more sophisticated approach to height estimation, e.g. by fitting a geometric primitive to the head's point cloud.

Common challenges in monocular analysis can be tackled with stereoscopic approaches. However, stereo vision is resource intensive and therefore reduced to single salient people in this work. Further reduction of the computational load could be achieved by incorporating knowledge about the expectable disparity range from matching results of directly preceding epochs.

Derived height estimates can be used to support subsequent monocular tracking. The person's position on the ground plane can be inferred from monocular observations of its head. This is particularly useful to overcome partial occlusion after the person left the overlapping fields of view.

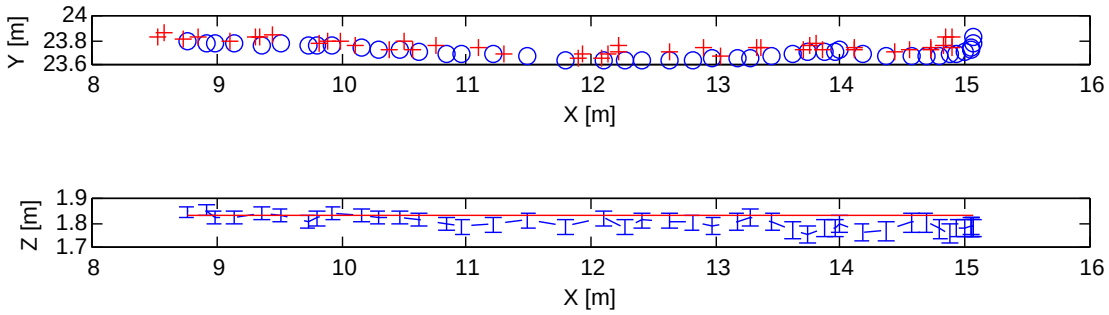

Figure 6: Results for a person approaching the cameras. Reference data is shown in red, our results in blue.

The presented work will be the basis of further investigations on the optimisation of matching costs and the analysis of the derived point cloud.

ACKNOWLEDGEMENTS

This research was funded by the German Federal Ministry of Education and Research (BMBF), 13N10809 - 13N10814.

REFERENCES

Haritaoglu, I., Harwood, D. and Davis, L., 1998. W^4S: A real-time system for detecting and tracking people in 2 1/2D. In: ECCV'98, LNCS, Vol. 1406, pp. 877–892.

Hirasaki, E., Moore, S. T., Raphan, T. and Cohen, B., 1999. Effects of walking velocity on vertical head and body movements during locomotion. Experimental Brain Research 127, pp. 117–130.

Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2), pp. 328 –341.

Jänen, U., Huy, M., Grenz, C., Hähner, J. and Hoffmann, M., 2011. Distributed three-dimensional camera alignment in highly-dynamical prioritized observation areas. In: Fifth ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC).

KaewTraKulPong, P. and Bowden, R., 2001. An improved adaptive background mixture model for real-time tracking with shadow detection. In: Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems.

Konolige, K., 1997. Small vision systems: Hardware and implementation. In: Proc. of 8th International Symposium on Robotics Research.

McGlone, J. C. (ed.), 2004. Manual of Photogrammetry. 5 edn, American Society for Photogrammetry and Remote Sensing.

Monari, E., Maerker, J. and Kroschel, K., 2009. A robust and efficient approach for human tracking in multi-camera systems. In: Proc. of Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009. AVSS '09, pp. 134 –139.

Zhao, T., Aggarwal, M., Kumar, R. and Sawhney, H., 2005. Real-time wide area multi-camera stereo tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, Vol. 1, pp. 976–983.

Zhou, J., Wan, D. and Wu, Y., 2010. The chameleon-like vision system. IEEE Signal Processing Magazine 27(5), pp. 91 –101.