INTEGRATION OF PRIOR KNOWLEDGE INTO DENSE IMAGE MATCHING FOR VIDEO SURVEILLANCE

M. Menze, C. Heipke

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
menze / heipke @ipi.uni-hannover.de

Commission III

KEY WORDS: Close Range, Stereoscopic Image Matching, Surface Reconstruction

ABSTRACT:

Three-dimensional information from dense image matching is a valuable input for a broad range of vision applications. While reliable approaches exist for dedicated stereo setups they do not easily generalize to more challenging camera configurations. In the context of video surveillance the typically large spatial extent of the region of interest and repetitive structures in the scene render the application of dense image matching a challenging task. In this paper we present an approach that derives strong prior knowledge from a planar approximation of the scene. This information is integrated into a graph-cut based image matching framework that treats the assignment of optimal disparity values as a labelling task. Introducing the planar prior heavily reduces ambiguities together with the search space and increases computational efficiency. The results provide a proof of concept of the proposed approach. It allows the reconstruction of dense point clouds in more general surveillance camera setups with wider stereo baselines.

1. INTRODUCTION

The automated analysis of surveillance videos is an important tool to support human operators. Given the enormous amount of data collected by omnipresent surveillance cameras manual on-line inspection of the images is impossible and even manual reconfiguration of pan-tilt-zoom cameras (PTZ-cameras) is very challenging. Often, human operators cannot prevent camera networks recording empty scenes while missing crucial events, a fact that obviously limits the efficiency of such systems. Pre-filtering of interesting scenes, automatic reconfiguration of cameras to focus relevant contents and the extraction of geometric information about people, actions and scene layout can help to make the task of manual inspection of video surveillance footage more tractable. Spatial information from stereoscopic analysis is very valuable in this context. However, low image resolution and wide baselines render the application of dense image matching for surveillance videos a challenging task. In common surveillance scenarios repetitive patterns and a significant spatial extent of the scene of interest demonstrate the limits of most general image matching approaches.

In this paper, we show how to derive strong prior knowledge for image matching assuming a realistic video surveillance setup and we describe the integration of this prior knowledge into a graph-cut based image matching framework. The method is proposed as a building-block for more comprehensive surveillance applications. While it tackles typical challenges to dense image matching and aims at reliable depth estimation, it does not address further tasks like people detection and tracking.

A brief overview of related work and the details of our method are given in section 2. Section 3 describes our experimental evaluation of the approach. The results provide a proof of concept and show the improved quality of the derived point clouds.

2. METHOD

2.1 Related work

Stereo image matching on dedicated short-baseline image pairs has been a major topic of research throughout the last decades. The taxonomy by (Scharstein et al. 2002) gives an insight into common dense matching methods. Among the best performing approaches are those enforcing global smoothness assumptions. Graph-cut based optimization strategies (Boykov et al. 2001) are employed for efficient inference of optimal disparities and are widely applied to diverse optimization tasks in photogrammetry and computer vision.

Because in related work on video surveillance stereoscopic image matching is often regarded as the central component in the respective systems, the used sensors or sensor networks are mostly designed to fulfil the specific 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 Darrel 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 are 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 to integrate strong prior knowledge into the process of dense image matching so that it becomes applicable for more general camera setups with wider stereo baselines.

2.2 Setup

The goal of the presented work is to provide a dense image matching approach that can be applied in realistic surveillance camera networks without dedicated stereo sensors. Camera calibration and absolute orientation are assumed to be known so that for camera pairs with sufficiently overlapping fields of view the images can be transformed to the normal case of

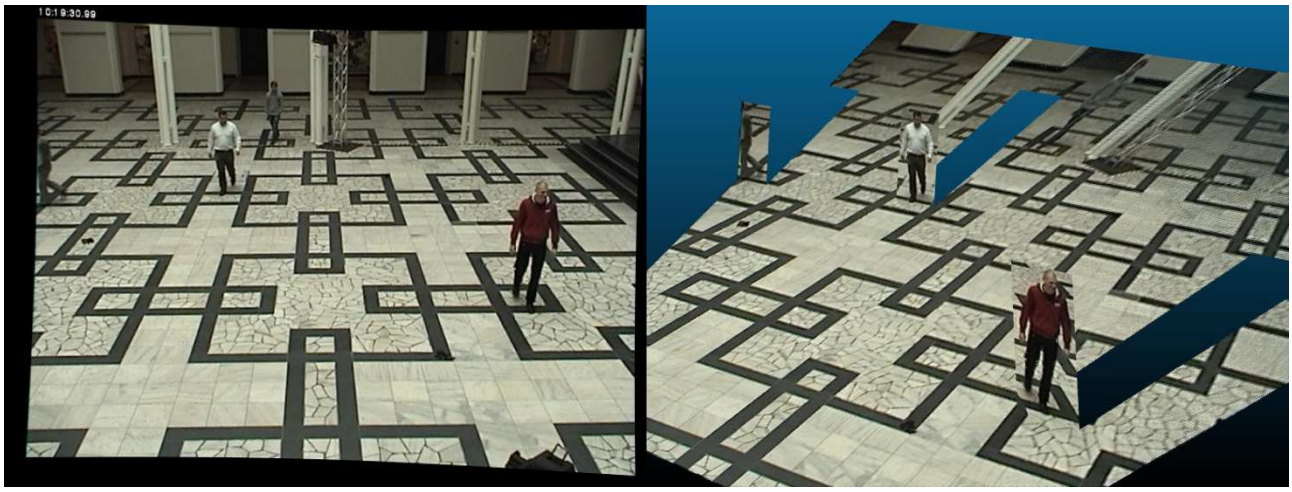

Figure 1. Rectified input image (left) and visualization of planar prior (right).

stereophotogrammetry. In addition, the exterior orientation is assumed to refer to an object coordinate system that depends on a predominantly horizontal ground plane so that strong prior knowledge can be derived for objects moving on this plane.

2.3 Derivation of prior knowledge

To extract moving objects we make use of a common pre-processing step and remove static background from the image sequences by subtracting an adaptive background model (Kaewtrakulpong et al. 2001). The resulting foreground blobs are enhanced by morphologic closing. They are used to instantiate planes that give a first approximation of the desired result in object space. To exclude small blobs induced by noise an area-threshold is applied. The handling of merging situations between multiple foreground objects is outside the scope of this paper. Certainly, depth cues from image matching would be valuable for tackling this kind of challenge.

In the context of video surveillance applications it is reasonable to assume that objects of interest predominantly extend perpendicular to the ground. Thus, we represent prior knowledge for the image matching approach as upright planes with 3D normals pointing parallel to the ground and perpendicular to the cameras' x axis. Given the exterior and interior camera orientation a first estimation of the position of the object on the ground plane can be derived from monoplotting, i.e. intersection of the viewing ray through the bottom-most point of the foreground blob with a plane in object space. Our straight-forward implementation directly uses the ground plane for this purpose. Shadows and occlusions may cause errors in this simple reconstruction but the resulting planes still give reasonable priors on the 3D positions of the observed objects. To support image matching these planes can be projected to disparity space, which is the discretization of the field of view in image coordinates and disparity. In surveillance applications the cameras are usually mounted above the volume of interest and tilted towards the ground so that upright planes in the object coordinate frame project to slanted planes in disparity space. Figure 1 depicts a typical input image rectified to epipolar geometry (on the left) and colour information from the image projected to the approximate 3D planes. Note that there is no plane instantiated for one person in the background since the size of the corresponding foreground blob is smaller than the respective threshold.

2.4 Disparity estimation

In order to derive a detailed disparity map of the observed object given the approximate plane, the disparity offset with respect to this plane has to be computed for each pixel of the respective foreground blob. This task can be formulated as a multi-class labelling problem.

A row-wise disparity seed point is directly specified by the plane in disparity space. The set of labels represents discrete disparity offsets in the range of feasible deviations from the planar prior. Admissible offsets are individually computed for each foreground object. Since they are defined in disparity space they depend on the absolute viewing distance. We found that for our application a metric search range of 1.5 m around the planar prior yields good results. It can be computed from the initial plane localization and transformed to disparity space at runtime. Note that this range is much smaller than the admissible range of disparities for the complete scene which leads to a massive reduction of ambiguities and a decreased computational burden. This also reduces the risk of the optimization getting stuck in local minima. Furthermore, the slope of the plane in disparity space would lead to varying search intervals along the vertical extent of the foreground blob when working with absolute disparity values. The use of disparity offsets with respect to the seed point circumvents this issue.

The task of finding optimal disparity offsets now corresponds to finding an optimal labelling of all foreground pixels. The relative quality of a labelling L is commonly evaluated by an appropriate energy functional of the form

$$E(L) = \sum_{p \in \boldsymbol{P}} E_D(l_p) + w_S \sum_{p,q \in \boldsymbol{N}} E_S(l_p, l_q) \qquad (1)$$

where $E(L)$ = energy induced by a labelling L
E_D, E_S = data, smoothness term

Equation 1 represents a Markov Random Field evaluating the labelling L by a weighted sum of a data term E_D and a smoothness term E_S with w_S controlling the influence of the smoothness term.

The data term measures the dissimilarity of pixels in the left and right stereo frame that are associated by the currently assigned

label or the corresponding absolute disparity. In our implementation the data term (2) is computed as the Hamming distance between the local binary pattern c_p (Zabih et al. 1994) around pixel p in the left image and the same descriptor from the right image at a horizontal offset induced by the label l_p of pixel p. By truncating the data cost term at τ_D it becomes robust against outliers. The choice of labels assigned to such pixels is dominated by the smoothness term (3).

$$E_D(l_p) = \min\left(\left|c_p^l - c^r(l_p)\right|, \tau_D\right) \quad (2)$$

The smoothness term E_S favours smooth transitions of disparity by penalizing the absolute difference of labels assigned to adjacent pixels. To allow for discontinuities, e.g. around the limbs, we use a robust energy function that evaluates the truncated absolute difference of the discrete labels in a 4-connected neighbourhood.

$$E_S(l_p, l_q) = \min\left(\left|l_q - l_p\right|, \tau_S\right) \quad (3)$$

Like the search range, the smoothness term is adjusted to the viewing distance. Obviously, discretization of the disparity space leads to significantly different ranges of plausible smooth transitions on object surfaces in different distances. In equation (3) this results in distance-dependent values of τ_S.We compute the truncation threshold from a quadratic function of the viewing distance.

Finding a globally optimum solution to such a multi-class labelling problem is in general NP-hard because of the complexity of the solution space. For appropriately defined energy terms graph-cut based approaches (Boykov et al. 2001) can be used to find approximate solutions. Starting from an initial labelling individual labels are iteratively expanded so that the total energy of equation (1) successively decreases.

3. RESULTS

The experiments are conducted on videos we collected for the joint research project CamInSens (CamInSens, 2013). The cameras are mounted 4.5 m above the ground plane with a stereo base of 4 m. Defining a region of interest of 10 m x 10 m in the scene the total range of plausible disparities for this setup spans approximately 250 pixels. Given the wide baseline and ambiguous image content standard matching algorithms fail to produce reliable dense results on these stereo pairs.

Because there is no densely labelled reference data we present qualitative results in figure 2 and validate the results on sparse, manually annotated control points on people in the scene. To this end, we annotated 10 equally spaced frames of our test sequence. Although a more extensive evaluation would be helpful, this validation ensures that there are no systematic errors in the approach and provides a proof of concept.

For the experiments we set the truncation threshold of the data term τ_D to 12, the truncation threshold for the smoothness term is individually set at runtime to enforce smooth disparity transitions corresponding to a 3D range of 0.8 m. The relative weight of data and smoothness term is set to 3. The energy given in equation (1) is minimized using graph-cuts and α–expansion (Boykov et al. 2001).

Table 1 gives the results of the quantitative evaluation at sparse control points. Resulting disparities are compared to horizontal offsets from manual annotation. To evaluate the results an inlier-threshold of 3 pixels is applied to the absolute difference of the disparity values. While the planar prior alone fails to predict accurate disparities, a standard winner-takes-all evaluation of the data term increases performance but still gives wrong results for more than half of the control points. The complete approach gives correct results for more than 75 % of the points providing a proof of concept and indicating remaining challenges. Figure 3 depicts one of the annotated frames with green crosses marking successfully matched control points and red crosses indicating outliers. The reconstructions on the right hand side of figure 3 show that outliers occur on the endpoints of limbs which are hard to match correctly due to the limited resolution of the input images and significant changes in perspective for the rightmost person.

	Prior	WTA	Graph-cut
% in	5.6	46.5	76.1
% out	94.4	53.5	23.9

Table 1. Percentage of disparity inliers and outliers at sparse control points.

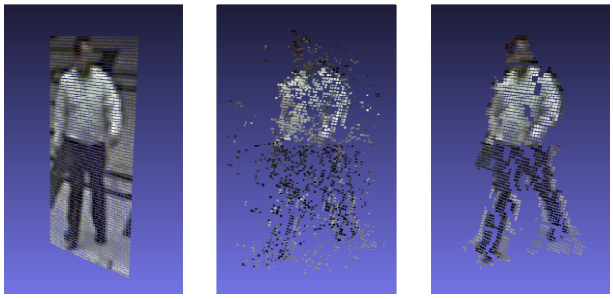

Figure 2. Planar prior (left), result of WTA (centre) and result of the proposed approach (right).

Figure 2 depicts an exemplary result of the optimization. The disparity maps are projected to point clouds in object space. The left part shows the planar prior, the centre depicts results of a simple winner-takes-all evaluation of the data term illustrating the challenging task. On the right the improved matching results are depicted yielding a consistent surface and a correct reconstruction of the limbs. Such results can directly be used for robust localization and estimation of body height and provide input to automated scene understanding.

4. CONCLUSIONS

We propose an approach addressing typical challenges to dense image matching in surveillance camera networks without dedicated stereo sensors. The integration of planar prior knowledge reduces ambiguities together with the search space and thus increases computational efficiency. While significantly improved results can be shown for isolated foreground blobs merging situations between multiple objects are not yet resolved. Future work will implement a feedback loop between matching and tracking to address this issue with the help of depth information. In a further step, a more detailed object model will be integrated to couple 3D reconstruction and semantic interpretation.

ACKNOWLEDGEMENTS

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

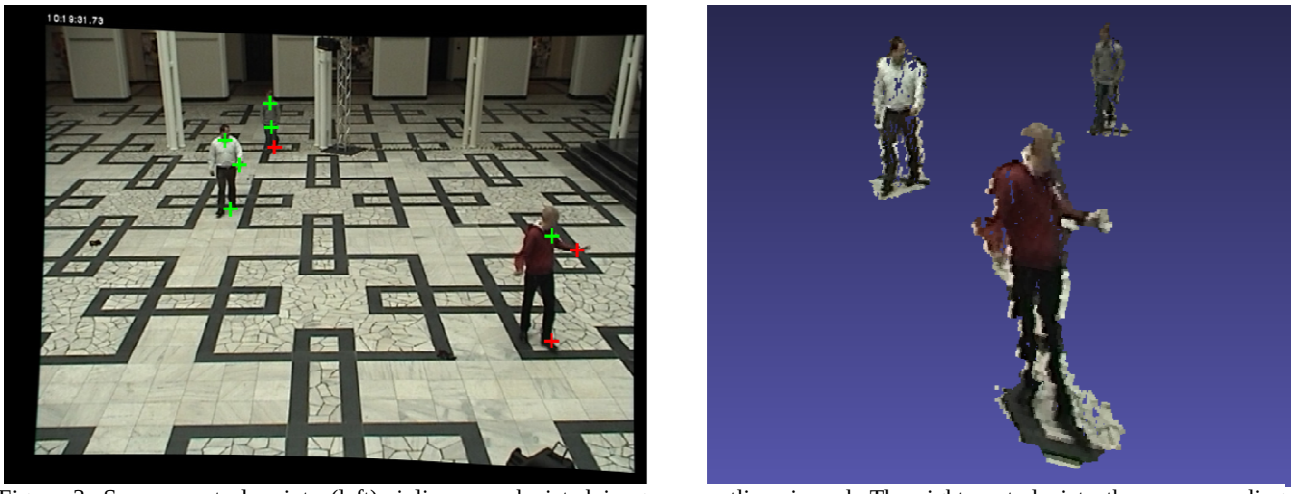

Figure 3. Sparse control points (left), inliers are depicted in green, outliers in red. The right part depicts the corresponding reconstructions.

REFERENCES

Boykov, Y., Veksler, O., & Zabih, R., 2001. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11), pp. 1222-1239

CamInSens, 2013. Project homepage. www.caminsens.org (12 July, 2014)

Haritaoglu, I., Harwood, D. & Davis, L., 1998. W4S: A Real-Time System for Detecting and Tracking People in 2.5 D, *Lecture Notes in Computer Science*, 1406, pp. 877–892

Kaewtrakulpong, P. & 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.*

Scharstein, D. & Szeliski, R., 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms, *International Journal of Computer Vision*, 47(1/2/3), pp. 7–42

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), pp. 523–537

Zabih, R. & Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: *Computer Vision - ECCV ´94*, Vol. II, pp. 151–158

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