THREE-DIMENISONAL MONOCULAR SCENE RECONSTRUCTION FOR SERVICE-ROBOTS
An Application

Sascha Jockel
CINACS, MIN Faculty, Dept. of Computer Science, University of Hamburg, Germany
jockel@informatik.uni-hamburg.de

Tim Baier-Löwenstein, Jianwei Zhang
Technical Aspects of Multimodal Systems, MIN Faculty, Dept. of Computer Science, University of Hamburg, Germany
{baier-loewenstein,zhang}@informatik.uni-hamburg.de

Keywords: 3D reconstruction - stereo vision - robotics - 3D model acquisition.

Abstract: This paper presents an image based three dimensional reconstruction system for service-robot applications in case of daily table scenarios. Image driven environment perception is one of the main research topics in the field of autonomous robot applications and fundamental for further action-plannings like three dimensional collision detection and prevention for grasping tasks.
Perception will be done at two spatial-temporal varying positions by a micro-head camera mounted on a six-degree-of-freedom robot-arm of our mobile service-robot TASER. The epipolar geometry and fundamentalmatrix will be computed by preliminary extracted corners of both input images detected by a Harris-corner-detector. The input images will be rectified using the fundamentalmatrix to align corresponding scanlines together on the same vertical image coordinates. Afterwards a stereo correspondence is accomplished by a fast Birchfield algorithm that provides a 2.5 dimensional depth map of the scene. Based on the depth map a three dimensional textured point-cloud is represented as interactive OpenGL scene model for further action-planning algorithms in three dimensional space.

1 INTRODUCTION

Three dimensional vision nowadays is a well investigated research field and was made easily accassible not only due to 3D pioneer Hartley (Hartley and Zisserman, 2003) or by work of Faugeras (Faugeras, 1993) and Ma et al. (Ma et al., 2004). Beyond car assistance systems, surveying, inspection and manufacturing techniques especially robotics can benefit from 3D vision. 3D environment maps can represent much more information than 2D plan views can do.

A variety of algorithms exists that each of them dealing with specific tasks of 3D vision. Tsai (Tsai, 1986) gives detailed explanation why camera calibration is fundamental to 3D sensing with robots. Moravec and Harris (Harris and Stephens, 1988) have done a lot of research in edge and corner detection which is fundamental for many correspondance analysis. To solve the epipolar geometry several approaches were proposed. Nonlinear methods as *Gauss-Newton* and *Levenberg-Marquardt optimization* from mathematics and *least-median-of-squares* (Rousseeuw and Leroy, 1987) and *RANSAC* (Fischler and Bolles, 1981). The *8-point algorithm* (Longuet-Higgins, 1981) is counted among the linear approaches. Trucco and Verri wrote an excellent overview of 3D vision (Trucco and Verri, 1998). Together with Fusiello they payed special attention to rectify image pairs (Fusiello et al., 2000), as also Pollefeys (Pollefeys et al., 2000), Hartley (Hartley, 1999) and Zhang (Zhang, 1998) did. The problem of stereo correspondance is approached in different ways with local windowing functions, graph cuts (Boykov et al., 1999) and fast dynamic algorithms by Birchfield and Tomasi with improved occlusion property (Birchfield and Tomasi, 1998).

This variety of specialized algorithms are well tested and documented but only several publications exists yet of cohere such task oriented algorithms to a single reconstruction system for real-life service-robot applications. Thus this work focus on the development of a three dimensional reconstruction system based on well-known and fast algorithms. Since the system is used with our service-robot TASER (Fig. 1)

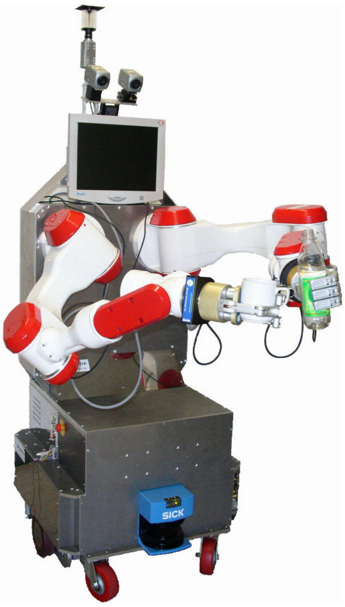

Figure 1: TASER service robot.

we have to pay attention to leave resources available for other indispensable robotic tasks, e.g. localisation, navigation, kinematics, sensor processing and motor control. Hence few subtasks of the reconstruction system uses algorithms supported by Intels OpenCV library (Intel, 2005) that fortunately are higly optimized to Intel core architectures. As input device a single mikro-head camera mounted on a six-degree-of-freedom robot-arm is used to obtain highest flexibility in the process of image acquisition concerning positioning and variability of baseline offset.

The remainder of this papers is organized as follows. In section 2 we briefly describe theoretical background to determine relative camera orientations and displacement. In section 3 we present the implementation of our reconstruction system for the mobile-robot system TASER. Experimental results are presented in section 4. Finally, we conclude in section 5 and give some remarks regarding the efficency of our implementation.

2 THEORETICAL BACKGROUND

Bevor describing the proposed three dimensional reconstruction system, we introduce some concepts of epipolar geometry.

Given a pair of views of a scene and a set of corresponding points x_i, x_i' in homogeneous coordinates, there exists a matrix $F \in \mathcal{R}^{3\times3}$, called the *fundamental matrix* (Faugeras, 1993), such that:

$$\mathbf{x}_i'^T F \mathbf{x}_i = 0 \;\; \forall i \qquad (1)$$

For any Point x_i (analogical x_i') in on view, the product Fp_i $(F^T p_i')$ defines a line, called the *epipolar line*, in the other view such that the corresponding point x_i' (x_i) belongs to this line. Moreover the right null vector of F (F^T) represents the *epipole* e (e') on the image plane. The fundamental matrix has maximum *rank* 2.

For image pairs the fundamental matrix F provides a constraint to identifying mismatches between image corners since corresponding corners are constrained to lie on respective epipolar lines.

An importand advantage to solve the correspondance problem is rectfication. Rectification determines a transformation of each image plane such that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes – usually the horizontal one (Fusiello et al., 2000). Thus computing the 2D stereo correspondance is reduced to a 1D correspondance search along the horizontal epipolar lines.

The resulting computaion of displacement of two corresponding pixels is called *disparity*. Together with the *baseline*, the distance between optical centers c and c' at both image capture position, the focal length and above-mentioned disparity via triangulation the 3D position of a pixel can be recovered.

3 IMPLEMENTATION

Fig. 2 illustrates the workflow of the developed three dimensional reconstruction system for our service-robot TASER. This robot is build completely upon of-the-shelf components without any custom products. The mobile plattform is equipped with a PA10-6C six-degree-of-freedom robot-arm from Mitsubishi Heavy Industries with an artificial BH-262 BarrettHand$^{\text{TM}}$mounted as tool with strain gauge sensor. Additionally a jAi micro-head camera is mounted at the BarrettHand$^{\text{TM}}$with an JK-L7.5M Toshiba lense as shown in Fig 3 (Baier et al., 2006). The robot is controled by a Pentium IV 2.4GHz standard computer.

In our system at early processing stages the images of both capture positions are processed separately and first will be merged for the computation of fundamental matrix. The processes of the flow-chart will be briefly described in the next paragraphs.

CAMERA CALIBRATION

The mikro-head hand-camera was calibrated a priori by a robust Tsai calibration algorithm (Tsai, 1986). The reason is as follows. Our mikro-head

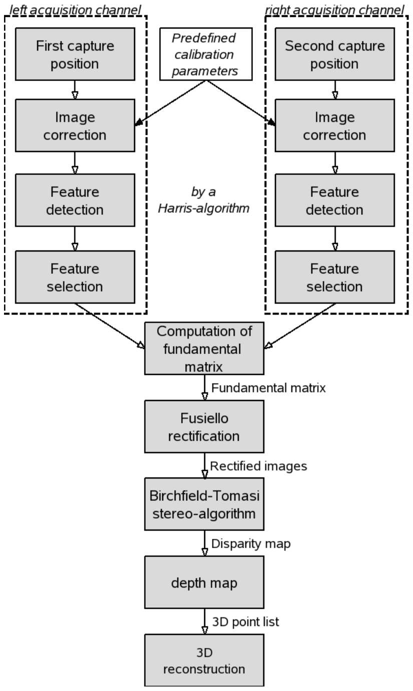

Figure 2: Simplified flow-chart of our 3D reconstruction system.

hand-camera is mounted on the manipulator in a rigid fixation tube (the white tube on the right of Fig. 3). We normally do not need to change the focus since we set it to infinity. Thus the predetermination of calibration coefficients is the easiest way to deal the next steps. For our purpose, it does not requiere to use structure from motion and online calibration techniques like Pollefeys presented in (Pollefeys, 1999). This process is shown in the flowchart only by its resulting calibration parameters.

IMAGE ACQUISITION AND CORRECTION

This paragraph describes the first two states of the reconstruction system together. The images of a daily table scenario are take by the mikro-head hand-camera mounted on the robot-arm. This setup enables a higly flexible image acquisition system within a cruising radius of approximatly 1000mm in all directions. Hence images can be aquired only by arm movements without moving the whole robot plattform. This is one of the main advantages of using a hand-camera.

Thereafter the distored input images are corrected using the predetermined calibration coefficient.

FEATURE EXTRACTION

This task is done with an algorithm by Harris. The Harris corner detector (Harris and Stephens, 1988) selects a pixel as corner if its responce R (see Eq. 2) is an 8-way maximum.

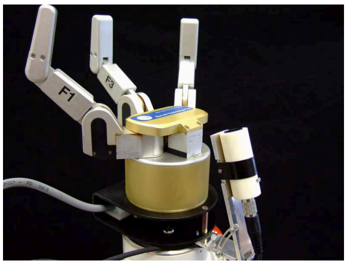

Figure 3: Micro-head camera (right) fix mounted at manipulator base.

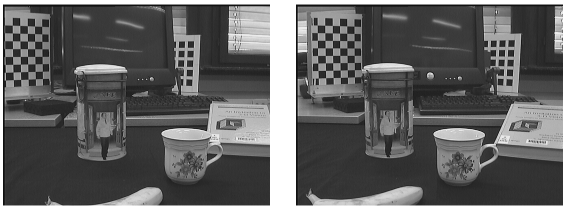

Figure 4: Input images at both camera position A (left) and B (right) with suitable orientation downwards and towards each other.

$$\mathbf{K} = det\mathbf{M} - k(trace\mathbf{M})^2 \tag{2}$$

$$\mathbf{M} = \begin{pmatrix} I_u^2 & I_{uv} \\ I_{uv} & I_v^2 \end{pmatrix} \tag{3}$$

where $det(\mathbf{M}) = I_u^2 I_v^2 - I_{uv}^2$ and $trace(\mathbf{M}) = I_u^2 + I_v^2$. Matrix $\mathbf{M}$ (3) is a covariance matrix with intensity values I for each pixel. Variable k is a weighting factor that was chosen to 0.04 by Harris. The image coordinates of extracted corners of both images then are passed to the next processing stages.

EPIPOLAR GEOMETRY

To compute the fundamental matrix one of three methods can be used. This options are possible due to the consequent software technical encapsulation. All other methods of the reconstruction system are exchangeable as well. The user can select wheter to use RANSAC - RANdom SAmpling Consensus - (Fischler and Bolles, 1981), LMedS - Least Median of Squares - (Rousseeuw and Leroy, 1987) or 8-point algorithm (Longuet-Higgins, 1981). By doing the latter the user has to select few corresponding points of the set of previously proposed corners.

For this variety of methods the Intel Open CV library is used. However, it figured out that the provided LMedS and RANSAC method does not work very reliable yet.

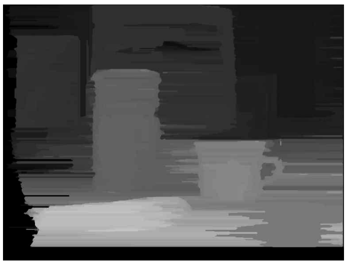

Figure 5: 2.5 dimensional depth map. The darker a pixel the farer away from camera.

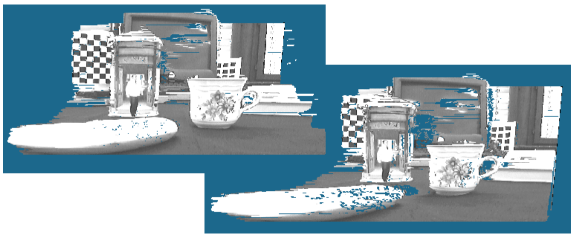

Figure 6: Two virtual views of the reconstructed 3D model based on precomputed disparity map.

RECTIFICATION

Fusiello et al. presented a robust, compact and easily reproducible algorithm that takes both perspective projection matrices of the original cameras to compute a pair of rectifying projection matrices (Fusiello et al., 2000). This matrices applied to the corrected input images leads to rectified views of our desk scene. The positions of the camera centers stays the same, whereas the orientation (the same for both cameras) differs from the old ones by suitable rotations. In contrast to Fusiello we are using one and the same camera at two spatial-temporal varying positions to acquire images without changing zoom and focus. Thus we use the same intrinsic camera matrix for both camera matrices described by Fusiello, only differ in their orientation.

STEREO ALGORITHM

Birchfield and Tomasi designed a fast and accurate stereo algorithm (Birchfield and Tomasi, 1996). It matches individual pixels in corresponding scanline pairs while allowing occluded pixels to remain unmatched. Espacially the latter is a main problem for conventional pixel and feature based algorithms, e.g. normalized cross-correlation, sum of squared differences and sum of absolute differences. Moreover Birchfields algorithm performs better results for homogeneous regions while using a measure of pixel dissimilarity that is insensitiv to image sampling. The disparity of corresponding pixels in two views is specified by the brightness of a pixel in a depth map (Fig. 5). The brighter a pixel the nearer to the camera.

Birchfields algorithm uses a cost function[1] (Eq. 4) that measures how unlikely it is that a sequence S describes the true correspondence.

$$\gamma(S) = N_{occ}\kappa_{occ} - N_s\kappa_r + \sum_{i=1}^{N_s} d(u_1,u_2) \tag{4}$$

where κ_r is a constant match reward of sequence s, κ_{occ} a constant occlusion penalty of s, N_{occ} and N_s are the number of occlusions and matches and $d(u_1,u_2)$ is the dissimilarity between pixels of both images at same scanline.

The stereo algorithm is not that time-consuming and computational expensive as classical methods that normally compares the similarity of frames of size $N \times N$ around pixels to determine their correspondance.

Birchfield provides a very fast stereo algorithm that deals large homogeneous regions very well too. These kind of regions often occure in environments of robots.

3D MODEL

Eq. 5 is used to compute the 3D coordinate of a pixel u, v. Therefore a pixel of the left image with known focal length through the calibration process is multiplied by a fraction of baseline b by disparity d.

$$\begin{pmatrix} x_c \\ y_c \\ z_c \end{pmatrix} = \frac{b}{d}\begin{pmatrix} u_l \\ v_l \\ f \end{pmatrix} \tag{5}$$

The disparity can be obtained from the depth map whereas the baseline is computed by the joint angles of the robot-arm. During acquisition process the arm pose is controlled via a homogeneous transformation $T = noap_{3\times4}$. The baseline for the second image is added to T by choosing $T_b = n_b o_b a_b p_{b3\times4}$. Whereas $n_b o_b a_b$ is set to identity $I_{3\times3}$ and $p_b = (0,y,0)^T$ as the corresponding baseline. The second position is then defined as $T' = TT_b$. Since we can translate the manipulator in camera coordinates (respectivelly to optical center) and the joints has a very well resolution we acquire good results for the baseline.

Fig. 6 shows the resulting three dimensional model. For displaying OpenGL is used and the 3D points are textured its original gray value of the left image.

[1] The cost function is justified solely by empirical evidence (Birchfield and Tomasi, 1996).

4 EXPERIMANTAL RESULTS

To validate the quality of our computation and to find a proper baseline we have done experiments. We placed six almost planar objects orthogonal at known distances to our camera setup. Then two images were recorded in a parallel orientation and vertical displacement. One of two input images for a test cycle with its resulting depth images is shown in Fig. 7. We repeated this experiment seven times with varying baselines to determine which baseline obtains best results in final reconstruction process for scenes from 800cm up to 2200cm. The results are shown in Fig. 8.

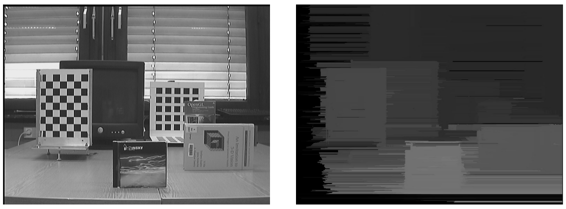

Figure 7: Left image of the experimental setup (left) and associated disparity map to determine optimal baseline offset.

The graph visualizes the relationship between real measured (abscissa) and computed distances by our reconstruction system (ordinate). As seen a small baseline leads to bad results. The reason is that the disparity resolution is very small and hence offers a rough depth reconstruction. If the baseline rises beyond 200mm the disparity resolution is appropriate but the depth images are cluttered. This happens due to problems with birchfields algorithm to find corresponding pixels at large disparities.

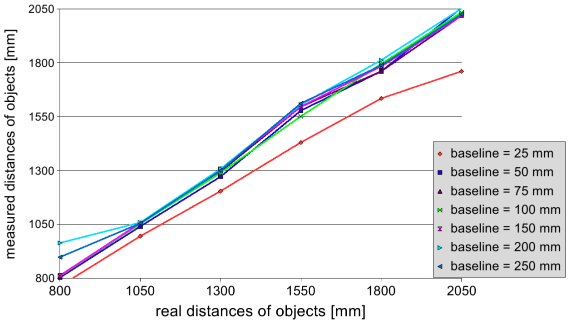

Figure 8: Comparison of real measured depth values (abscissa) to the measured results (ordinate) of our reconstruction system.

Within a baseline of 75mm til 150mm subjectively the images are bearable concerning cluttering and disparity resolution. Hence we choose a baseline of 100mm as suitable for the reconstruction task of desk scenes.

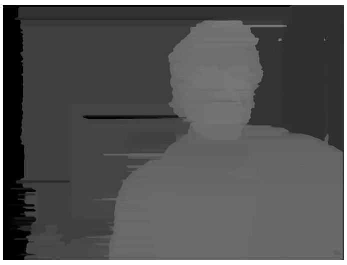

Figure 9: 2.5 dimensional depth map of a face. The darker a pixel the farer away from camera.

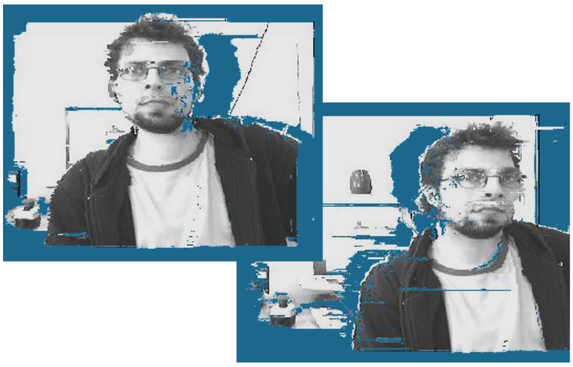

Figure 10: Two virtual views of a reconstructed face based on the disparity map of Fig. 9.

Moreover, tests shown that our reconstruction systems is able to deal with large laboratory rooms. But the baseline has to be larger than for desk scenarios. Unfortunately the results are not as well as for the developed desk reconstruction tasks. Furthermore it can be used to do proper 3D face reconstructions. Results for a tentatively face reconstruction are shown in Fig. 9 & 10.

5 CONCLUSION AND FUTURE WORK

By the use of a hand-camera we achieved a very flexible system for free image acqusition without moving the robot plattform. To vary the baseline during acquisition is manageable easily. This is a big advantage in small and tight environments to adapt to prevailing circumstances easily. Furthermore no robot components will cause occlusions as it will be with the stereo-rig at the top of TASER.

The usage of Open CV supported and optimized algorithms yields to good performance and leave sufficient resources available for other basic robotic tasks. Birchfields dynamic stereo correspondance is an appropriate solution concerning the results to its performance.

Adding more viewpoints to capture a real world scenario will lead to improved three dimensional models and will help reduceing occluded regions. Determining the *next-bext-view* described by Chen (Chen and Li, 2004) will keep the number of images needed for adequate reconstruction as small as possible. Based on full three dimensional models we will proceed with collision detection algorithms for robotarm interaction, e.g. grasping and manipulation, in three dimensional space. Furthermore we want to merge the reconstruction system with a computation of optimal object grasps presented by Baier (Baier and Zhang, 2006) .

REFERENCES

Baier, T., Hueser, M., Westhoff, D., and Zhang, J. (2006). A flexible software architecture for multi-modal service robots. In *Multiconference on Computational Engineering in Systems Applications (CESA)*.

Baier, T. and Zhang, J. (2006). Reusability-based semantics for grasp evaluation in context of service robotics. In *IEEE International Conference on Robotics and Biomimetics (ROBIO 2006)*, Kunming, China.

Birchfield, S. and Tomasi, C. (1996). Depth discontinuities by pixel-to-pixel stereo. Technical report STAN-CS-TR-96-1573, Stanford University.

Birchfield, S. and Tomasi, C. (1998). Depth discontinuities by pixel-to-pixel stereo. In *Proceedings of the Sixth International Conference on Computer Vision*, pages 1073–1080, Bombay, India.

Boykov, Y., Veksler, O., and Zabih, R. (1999). Fast approximate energy minimization via graph cuts. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 377–384.

Chen, S. Y. and Li, Y. F. (2004). Automatic sensor placement for model-based robot vision. In *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, volume 34, pages 393–408. IEEE Systems, Man, and Cybernetics Society.

Faugeras, O. (1993). *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, Cambridge, MA, USA.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, 24(6):381–395.

Fusiello, A., Trucco, E., and Verri, A. (2000). A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Fourth Alvey Vision Conference*, pages 147–151.

Hartley, R. I. (1999). Theory and practice of projective rectification. volume 35, pages 115–127. Kluwer Academic Publishers, Hingham, MA, USA.

Hartley, R. I. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Camebridge University Press.

Intel, C. (2005). Open CV 0.9.7. http://www.intel.com/research/mrl/research/opencv/.

Longuet-Higgins, H.-C. (1981). A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135.

Ma, Y., Soatto, S., Kosecka, J., and Sastry, S. S. (2004). *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer, Berlin, Heidelberg.

Pollefeys, M. (1999). *Self-Calibration and Metric 3D Reconstruction from Uncalibrated Image Sequences*. Ph.d. thesis, ESAT-PSI, Katholieke Universiteit Leuven.

Pollefeys, M., Koch, R., Vergauwen, M., and Gool, L. V. (2000). Automated reconstruction of 3d scenes from sequences of images. *ISPRS Journal Of Photogrammetry And Remote Sensing*, 55:251–267.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley, New York.

Trucco, E. and Verri, A. (1998). *Introductory Techniques for 3–D Computer Vision*. Prentice Hall, New York.

Tsai, R. Y. (1986). An efficient and accurate camera calibration technique for 3d machine vision. In *International Conference on Computer Vision and Pattern Recognition*, pages 364–374, Miami Beach, Fla. IEEE, IEEE Computer Society Press.

Zhang, Z. (1998). A flexible new technique for camera calibration. Technical report MSR-TR-98-71, Microsoft Research.