In W. Burger and J. Scharinger, editors, Digital Imaging in Media and Education
volume 179 of Schriftenreihe, pages 141-148. OCG, 2004. Proceedings of the 28th ÖAGM/AAPR Conference.

Sparse 3D Reconstruction of a Room

Gerald Schweighofer[1] Miguel Ribo[2] Axel Pinz[1]

[1]Institute of Electrical Measurement and Measurement Signal Processing

[2]Christian Doppler Laboratory for Automotive Measurement Research

Graz University of Technology, Schiesstattg.14B, A-8010 Graz, Austria

Abstract:
Nowadays many known ways exist to compute 3D models of office rooms. But all of them have one disadvantage: the long computation time. Within this work, the computation of a sparse 3D model of a room is described which is done in a short time. To do this we extend the stereo reconstruction method on panoramic images, and we calibrate the camera with a method which takes benefits from the image acquisition process, and from the innovative design of our self developed artificial targets. Experimental results show the feasibility of the proposed 3D reconstruction.

1 Introduction

The relation between a point $\mathbf{M}$ in 3D space and its corresponding point $\mathbf{m}$ in image plane is given by [6]

$$s\mathbf{m} = \underbrace{\mathbf{K}[\mathbf{R}|\mathbf{t}]}_{\mathbf{P}}\mathbf{M}, \tag{1}$$

where s is an arbitrary scale and $\mathbf{P}$ is a 3×4 matrix known as the *perspective projection* matrix. The points $\mathbf{m} = [u, v, 1]^T$ and $\mathbf{M} = [x, y, z, 1]^T$ are expressed in *homogeneous coordinates*. The 3×3 rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ describe the pose (i.e. the *extrinsic* parameters) of the camera with respect to the world coordinate system. The 3×3 matrix $\mathbf{K}$ contains the *intrinsic* parameters of the camera, such that

$$\mathbf{K} = \begin{bmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

where α_u and α_v are the scale factors in the $\mathbf{u}$ and $\mathbf{v}$ axis directions, respectively. The *principal point* (u_0, v_0) is the point where the optical axis of the camera intersects the image plane (we assume zero skew).

A closer look at equation (1) points out the problems that we need to address in order to compute a sparse 3D model of an office-like room. Indeed, as the camera used to reconstruct the scene is uncalibrated (i.e. $\mathbf{K}$ is unknown) we need to deal with the problem of camera self-calibration [5].

Moreover, as both the scene and the camera motion are unknown (i.e. $\mathbf{M}$ and $(\mathbf{R}, \mathbf{t})$) we need to deal with the problem of structure from motion [3].

Several types of methods to resolve (1) have been proposed in the literature. These methods can be categorised with respect to the number of images used to reconstruct the scene. In a single view method [4, 9] the reconstruction of a 3D structure can be done if vanishing points and lines can be computed from the image. For the stereo view approaches, we can use either the small baseline stereo method [14], the wide baseline stereo method [13], or a combination of both methods. Multiple views (captured by one or more cameras) methods are also proposed [1, 8, 9]. For those cases, the camera moves "randomly" in the scene. When the camera only rotates around the axis perpendicular to its optical axis, we are talking about panoramic view methods [12].
In this paper, we propose a multiple stereo views method with a zoom camera. The images are acquired in the way that we are able to build panoramic views of the scene.

2 Image acquisition process

For the acquisition of the scene's panoramic view, we use a zoom camera mounted on top of a Pan-Tilt-Unit (PTU). From the technical features of the camera, we can compute the maximum horizontal camera's field of view by

$$B_{\text{max}}^{\text{cam}} = 2 \arctan\left(\frac{x_{\text{max}}}{2f}\right) \tag{3}$$

where x_{max} is the maximum resolution of the camera in the x-direction, and f is the camera's focal length. In addition, if we consider that two consecutively acquired images should overlap each other of about α degrees, we define

$$I_\alpha = \text{ceil}\left(\frac{B_{\text{max}}^{\text{cam}} - \alpha}{B_{\text{max}}^{\text{PTU}}}\right) \quad \text{and} \quad A_\alpha = \frac{I_\alpha}{B_{\text{max}}^{\text{PTU}}} \tag{4}$$

where the parameter A_α indicates the rotation angle between two consecutive image views.

Figure 1 depicts the image planes of an acquisition process while the PTU is rotating about A_α degrees. The PTU rotates to its leftmost side position, and then the first image is captured (image 1). The PTU gradually rotates A_α degrees to the right, and at every step the camera captures a new image. The image acquisition process is ended when the PTU reaches its rightmost side position.
As a result, we are able to capture a panoramic view of the scene. We remark that the overlap area between the first and the last image is smaller than for the other images (images 1 and 13 in fig. 1), because the maximum horizontal rotation of the PTU is $317°$.

Our method requires to take two panoramas of this kind (the tripod with both PTU and camera has to be moved to an arbitrary second position) in order to build stereo panoramic images (the stereo baseline is the distance between the two tripod positions).

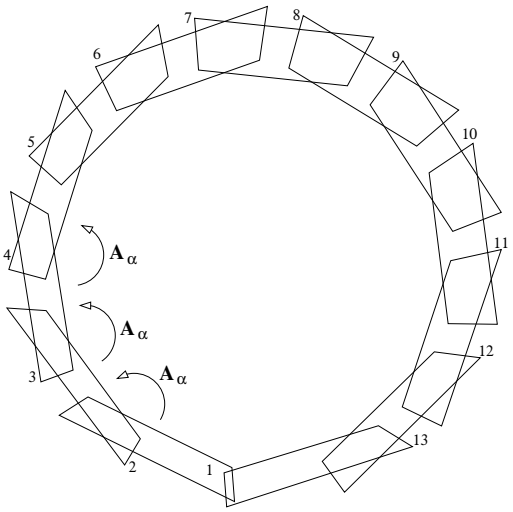


Figure 1: Acquisition of a panoramic view.

3 Calibration

Camera calibration means the process from which we are able to compute both the *intrinsic* (i.e. internal geometry) and the *extrinsic* (i.e. external geometry) parameters of the camera [6].

Intrinsic parameters: As we use a zoom camera (i.e. variable focal length), the intrinsic parameters of the camera need to be computed before the 3D reconstruction process starts. On the assumption that the PTU's axis of rotation goes through the camera's centre of projection, we can rewrite the equation (1) as

$$s\mathbf{m}_{ij} = \mathbf{K}\mathbf{R}_j\mathbf{M}_i \tag{5}$$

where $\mathbf{m}_{ij}$ is the projection of the scene point $\mathbf{M}_i$ on the image j. In other words, we can say that the point $\mathbf{M}_i$ must belong to the straight line

$$\mathbf{l}_i = (\mathbf{K}\mathbf{R}_j)^{-1}\mathbf{m}_{ij} \tag{6}$$

Moreover, as explained in [5], the relationship between two different images can be written as

$$s\mathbf{m}_{ik} = \mathbf{K}\mathbf{R}_k\mathbf{M}_i = \mathbf{K}\mathbf{R}_k(\mathbf{K}\mathbf{R}_j)^{-1}\mathbf{m}_{ij} = \mathbf{H}_{kj}\mathbf{m}_{ij} \tag{7}$$

where $\mathbf{H}_{kj}$ is a 3×3 non-singular homogeneous matrix. This matrix is a *homography* [6] and relates the analytical correspondence between points of images j and k, respectively. On the assumption that $\mathbf{R}_j = [[1,0,0]^T, [0,1,0]^T, [0,0,1]^T]$ (w.r.t. the second image k), equation (7) gives after some algebra

$$\mathbf{K}\mathbf{R}_k - \mathbf{H}_{kj}\mathbf{K} = 0 \tag{8}$$

This equation gives us the way of how to compute the intrinsic parameters of the camera if at least four point correspondences are found in both images j and k. To solve (8), we favour to use the RANSAC algorithm [7]. Moreover, since the camera's centre of projection does not exactly belong

to the PTU's axis of rotation, we apply also a non-linear optimisation process in order to improve the accuracy of the computed intrinsic parameters.

Extrinsic parameters: Algebraically, if $(\mathbf{R}, \mathbf{t})$ defines the rigid displacement from the image 1 to the image 2 camera coordinate systems, we have the relation

$$\mathbf{v}_2 = \mathbf{R}\mathbf{v}_1 + \mathbf{t} \tag{9}$$

where $\mathbf{v}_i$ is the vector defined by equation (6) as $\mathbf{l}_i/\|\mathbf{l}_i\|$. Taking the vector product with $\mathbf{t}$, followed by the scalar product with $\mathbf{v}_2$ we have

$$\mathbf{v}_2 \cdot (\mathbf{t} \times \mathbf{R}\mathbf{v}_1) = 0 \tag{10}$$

which expresses that the vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{t}$ are coplanar. By rewriting (10) as $\mathbf{v}_2^T \mathbf{E} \mathbf{v}_1 = 0$, we can derive the so-called 3×3 *essential* matrix

$$\mathbf{E} = [t]_\times \mathbf{R} \text{ where } [t]_\times = \begin{pmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{pmatrix} \tag{11}$$

The computation of the matrix $\mathbf{E}$ is done using the "8-point algorithm" [11]. Then, by means of the *singular value decomposition* method, we are able to compute the translation $\mathbf{t}$ and the rotation matrix $\mathbf{R}$. Finally, using the intrinsic matrix $\mathbf{K}$ to relate points in the camera coordinate system to image points in pixels, we have

$$\mathbf{m}_2^T \underbrace{\mathbf{K}^{-T}\mathbf{E}\mathbf{K}^{-1}}_{\mathbf{F}} \mathbf{m}_1 = 0 \tag{12}$$

where $\mathbf{F}$ is the 3×3 *fundamental* matrix which defines the *epipolar geometry* [6] needed to find corresponding image points between two distinct camera's positions.

4 Landmarks and 3D reconstruction

To compute the extrinsic parameters of the camera (see section 3), we need to accurately find several corresponding image points between various camera positions. For this purpose, we have designed unique targets (see fig. 2a) which are positioned in the scene (prior to the image acquisition process) at unknown locations. Under perspective projection, the invariant properties of such targets are the intersection points of the tangents between two circles. We remark that for every target the two bottom circles (see fig. 2a) are identical, while the size of the top circles is used to identify the target itself (useful to find corresponding targets among image pairs). As a result, by using an iterative method, we are able to extract (from four tangents) eight projective invariant points for each pair of ellipses (see fig.2b). A similar idea to extract the extrinsic parameters from two circles is described in [2].

While artificial targets are used for the extrinsic parameter estimation, the sparse 3D reconstruction of the scene is based on natural landmarks (i.e. corner based features). These points are extracted

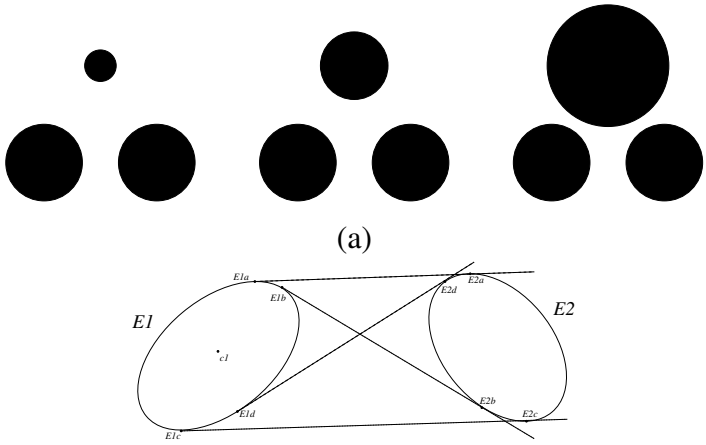


Figure 2: (a) Example of our targets. (b) Under perspective projection, four tangents between two ellipses give eight invariant points.

among all images using the method proposed by Harris and Stephens [10]. Afterwards, by means of the epipolar constraint (12) corresponding points can be found, and the 3D reconstruction can be performed.

Reconstruction is here the process of computing three-dimensional structure of the scene from two-dimensional image information (e.g. point based features). Given two image points $\mathbf{m}_1$ and $\mathbf{m}_2$ from two distinct camera positions, we can compute by means of the projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$

$$s\mathbf{m}_1 = \mathbf{P}_1\mathbf{M} \quad \text{and} \quad s\mathbf{m}_2 = \mathbf{P}_2\mathbf{M} \tag{13}$$

where $\mathbf{M}$ is a 3D point of the scene. Then, equation (13) can be written in the form (for a suitable 4×4 matrix $\mathbf{B}$)

$$\mathbf{BM} = 0 \quad \text{with} \quad \mathbf{B} = \begin{bmatrix} \mathbf{p}_1^1 - u_1\mathbf{p}_1^3 \\ \mathbf{p}_1^2 - v_1\mathbf{p}_1^3 \\ \mathbf{p}_2^1 - u_2\mathbf{p}_2^3 \\ \mathbf{p}_2^2 - v_2\mathbf{p}_2^3 \end{bmatrix}, \tag{14}$$

where $\mathbf{p}_1^i$ and $\mathbf{p}_2^i$ are the i-th *row vectors* of the matrices $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively. Since this equation defines the point $\mathbf{M}$ up to a scale factor, we can impose the constraint $\|\mathbf{M}\| = 1$. The solution of equation (14) is simply the eigenvector of the matrix $\mathbf{B}^t\mathbf{B}$ associated to the smallest eigenvalue. This stage can be done again by using the singular value decomposition method.

5 Experimental results

For image acquisition we have used a Sony firewire zoom camera (DFW-VL-500) mounted on top of a pan-tilt-unit. Five artificial targets (see fig. 2a) were fixed around the room. The size of the room is approximately 4.5 x 6.0 m. We note that we do not need to measure any distance or relative positions between any of the targets. The camera was positioned at two distinct locations to acquire the panoramic views of the office room. After choosing the desired focal length of the camera (this

may vary w.r.t. the size of the room), we carry out the sparse 3D reconstruction of the room as described in fig. 3. The right hand side of fig. 5 shows the 3D reconstruction of the five targets used to calibrate the camera. Assuming perpendicular walls of the room, we get a mean angle error for the five targets of $4.14°$. The left hand side of fig. 5 shows the sparse 3D model of the room after the reconstruction of up to 161 points as explained in section 4. We remark that the model contains some outliers (9 points) which are essentially due to mismatch during the correspondence process. For the evaluation of the accuracy of the presented method, several gauge targets (i.e. strap-like

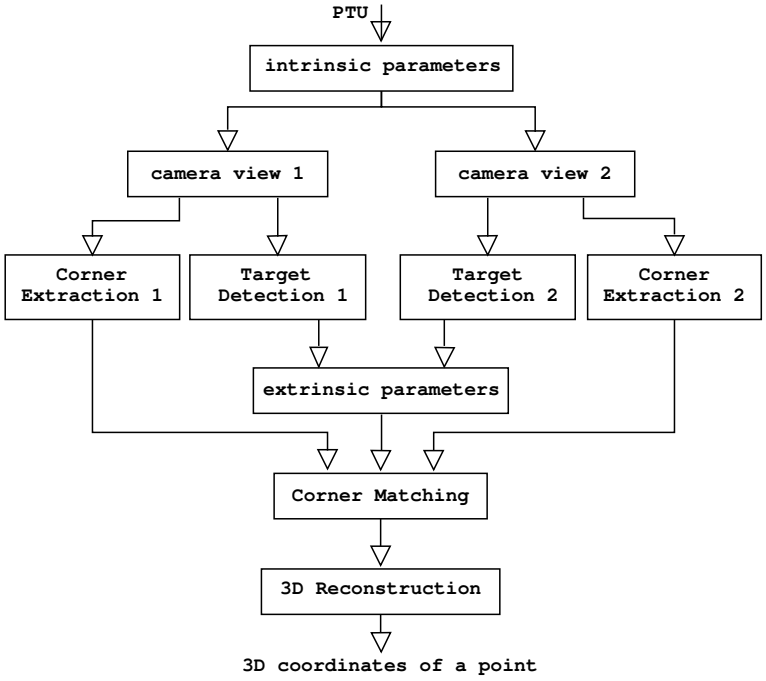


Figure 3: 3D reconstruction block diagram.

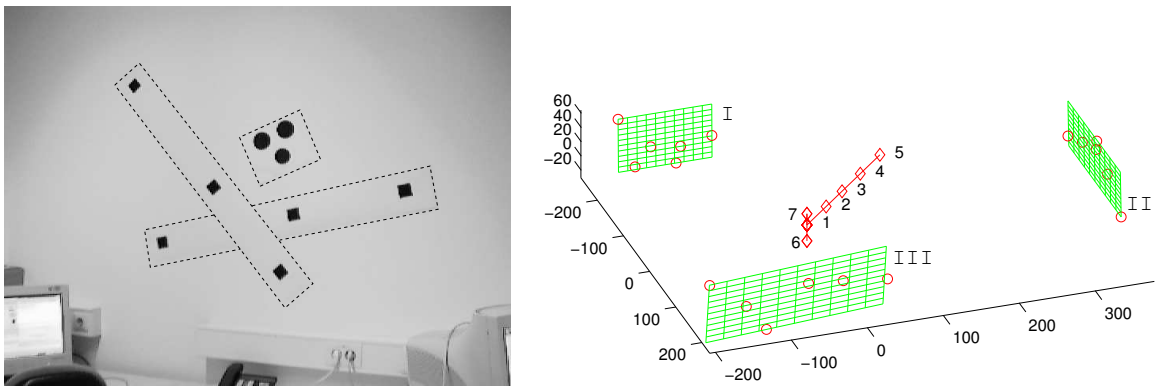
format with known length) are fixed on the walls of the room. Figure 4a shows such a constellation where two gauge targets are located near a calibration target. Figure 4b shows the 3D reconstruction of three planes (noted I, II, and III) each of them defined by two gauge targets. The diamond-styled markers show the locations of the camera where the panoramic views are taken. Table 1 describes the parameters coming from four (a, b, c and d) experimental evaluation cases. *Cam. pos.* are the

case	Cam. pos.	B (cm)	Pts	Tgs	nb. of matches	$\bar{l}$ (cm)	$\bar{d}$ (cm)
a	1-2	54.7	161	5	5968	5.03	2.48
b	1-3	99.6	146	5	7291	4.38	1.06
c	1-4	151.6	106	4	6930	3.80	1.48
d	1-5	207.2	105	5	6765	8.63	1.84

Table 1: Experimental evaluation data.

camera positions as plotted in fig. 4b. The value *B* indicates the stereo baseline for the related case. The value *Pts* indicates the number of points for which we compute a 3D reconstruction. The value

Tgs indicates how many calibration targets we found. The *number of matches* indicates how many correspondences we found. In addition (related to each case), the value $\bar{l}$ indicates the mean error concerning the length (i.e. about 1 m) of all reconstructed gauge targets, and the value $\bar{d}$ indicates the mean distance of all reconstructed black squares (see fig. 4a) w.r.t. the planes defined by two gauge targets.



(a) (b)

Figure 4: (a) Gauge targets used for the accuracy evaluation, and (b) their 3D reconstruction. The diamonds depict camera locations.

6 Conclusions

We have presented and described a method for the sparse 3D reconstruction of an unknown office room. To do that, we have extended the stereo reconstruction approach on panoramic images. Our method tackles two main problems which are *camera self-calibration* and *structure from motion*. The emphasis of the proposed method is also laid on flexibility (i.e. independent of the room's size), execution time (i.e. less than 30 min in matlab), and robustness. The experimental results show the feasibility of the proposed method, and the quality/accuracy of the computed 3D model.

Acknowledgements

This work was supported by the European project VAMPIRE - Visual Active Memory Processes and Interactive REtrieval (IST-2001-34401), the Austrian Science Foundation (FWF, project S9103-N04) and by the Christian Doppler Laboratory for Automotive Measurement Research.

References

[1] Q. Chen and G. Medioni. Efficient iterative solution to m-view projective reconstruction problem. In *IEEE CVPR*, volume 1, pages 23–25, Ft. Collins, USA, June 1999.

[2] Q. Chen, H. Wu, and T. Wada. Camera calibration with two arbitrary coplanar circles. In *European Conference on Computer Vision*, pages 521–532, May 2004.

[3] A. Chiuso, R. Brockett, and S. Soatto. Optimal structure from motion: Local ambiguities and global estimates. *International Journal of Computer Vision*, 39(3):195–228, Sep.-Oct. 2000.

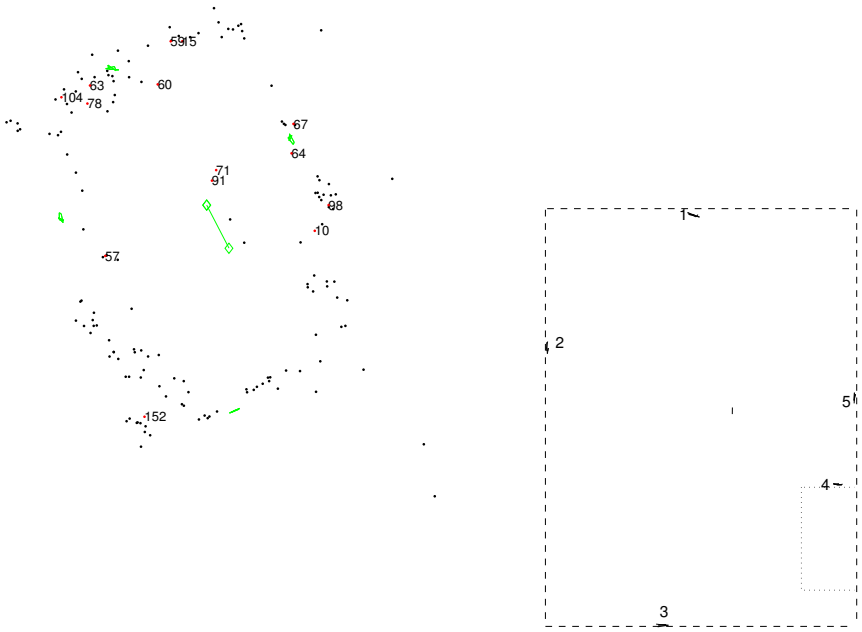


Figure 5: Top view of the sparse 3D reconstruction (left), and the room with the computed calibration target locations (right).

[4] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology". *International Journal of Computer Vision*, 40(2):123–148, November 2000.

[5] L. de Agapito, E. Hayman, and I. Reid. Self-calibration of rotating and zooming cameras. *International Journal of Computer Vision*, 45(2):07–127, November 2001.

[6] O. Faugeras. *Three-Dimensional Computer Vision: a Geometric Viewpoint*. MIT Press, 1993.

[7] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.

[8] A. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *IEEE CVPR*, volume 1, pages 125–132, Kauai, Hawaii USA, December 2001.

[9] E. Grossmann, D. Ortin, and J. Santos-Victor. Algebraic aspects of reconstruction of structured scenes from one or more views. In *British Machine Vision Conference*, pages 633–642, Manchester, England, September 2001.

[10] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of the* 4^{th} *Alvey Vision Conference*, pages 189–192, Manchester, 1988.

[11] R. Hartley. In defence of the 8-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, June 1999.

[12] P. Peer and F. Solina. Panoramic depth imaging: Single standard camera approach,. *International Journal of Computer Vision*, 47(1):149–160, April 2002.

[13] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *IEEE International Conference on Computer Vision*, pages 754–760, Bombay, India, January 1998.

[14] S. Shah and J. Aggarwal. Mobile robot navigation and scene modeling using stereo fish-eye lens system. *Machine Vision and Applications*, 10(4):159–173, 1997.