A Geometric Approach to Lightfield Calibration

R. Koch[1], B. Heigl[2], M. Pollefeys[1], L. Van Gool[1], and H. Niemann[2]

[1] Center for Processing of Speech and Images (PSI), K.U.Leuven, Belgium
[2] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
email: Reinhard.Koch@esat.kuleuven.ac.be, heigl@informatik.uni-erlangen.de

Abstract. Lightfield rendering allows fast visualization of complex scenes by view interpolation from images of densely spaced camera viewpoints. The lightfield data structure requires calibrated viewpoints, and rendering quality can be improved substantially when local scene depth is known for each viewpoint. In this contribution we propose to combine lightfield rendering with a geometry-based structure-from-motion approach that computes camera calibration and local depth estimates. The advantage of the combined approach w.r.t. a pure geometric structure recovery is that the estimated geometry need not be globally consistent but is updated locally depending on the rendering viewpoint. We concentrate on the viewpoint calibration that is computed directly from the image data by tracking image feature points. Ground-truth experiments on real lightfield sequences confirm the quality of calibration.

1 Introduction

There is an ongoing debate in the computer vision and graphics community between geometry-based and image-based scene reconstruction and visualization methods. Both methods aim at realistic and fast rendering of 3D scenes from image sequences.

Geometric reconstruction approaches generate explicit 3D scene descriptions with polygonal (triangular) surface meshes. A limited set of camera views of the scene is sufficient to reconstruct the 3D scene. Texture mapping adds the necessary fidelity for photo-realistic rendering to the object surface.

Image-based rendering approaches like lightfield rendering [14] and the lumigraph [6] have lately received a lot of attention, since they can capture the appearance of a 3D scene from images only, without the explicit use of 3D geometry. Thus one may be able handle scenes with complex geometry and surface reflections that can not be modeled otherwise. Basically one caches all possible views of the scene and retrieves them during view rendering.

Both approaches have their distinct advantages and weak points. In this contribution we discuss the combination of image-based rendering with a geometric structure-from-motion approach to obtain lightfields from image sequences of a freely moving camera. The necessary camera calibration and local depth estimates are obtained with the structure-from-motion approach. We will first give a brief overview of image-based rendering and geometric reconstruction techniques. We will then focus on the calibration problem for lightfield acquisition

from hand-held camera sequences. Experiments on lightfield calibration and geometric approximation conclude this contribution.

2 Image-based rendering

Image-based rendering techniques allow to capture a scene with a principally unlimited geometric complexity, with complex lighting and specular surface reflections. The view rendering depends only on the efficiency of data access and not on the scene complexity, hence rendering in constant time is possible. The price to pay for this advantage is a very high amount of data and a tedious image acquisition. In fact, one has to obtain the plenoptic function of the scene space with viewing rays in all possible positions, which is a 5-dimensional function. Perfect rendering is possible only if all viewing rays of a newly rendered view intersect the focal centers of originally acquired views. Interpolation between viewpoints will cause a distortion that is dependent on the scene geometry as well. The amount of views to be acquired is limited by the storage requirements, since a dense view sampling of a scene might easily generate Gigabytes of image data. Therefore one must try to compress the data efficiently by removing the inherent redundancy. Since the approach is strictly image-based, no viewpoint extrapolation is possible. Furthermore the geometry is encoded implicitly in the data and there is no way to change geometric scene properties e.g. for animations.

Recently two equivalent realizations of the plenoptic function were proposed in form of the lightfield [14], and the lumigraph [6]. They handle the case when we observe an object surface in free space, hence the plenoptic function is reduced to four dimensions (light rays are emitted from the 2-dimensional surface in all possible directions). The 4-D lightfield data structure employs a two-plane parameterization (see fig. 1). Each light ray passes through two parallel planes with plane coordinates (s, t) and (u, v). Thus the ray is uniquely described by the 4-tuple (u, v, s, t). The (s, t)-plane is the *viewpoint plane* in which all camera focal points are placed on regular grid points. The (u, v)-plane is the *focal plane* where all camera image planes are placed with regular pixel spacing. The optical axes of all cameras are perpendicular to the planes. This data structure covers one side of an object. For a full lightfield we would need to construct six such data structures on a cube around the object.

New views can be rendered from this data structure by placing a virtual camera on an arbitrary viewpoint and intersecting the viewing ray r with the two planes at (s, t, u, v). The resulting radiance is a simple radiance lookup for r. This, however, applies only if the viewing ray passes through original camera viewpoints and pixel positions. For rays passing in between the (s, t) and (u, v) grid coordinates an interpolation is applied that will degrade the rendering quality depending on the scene geometry. In fact, the lightfield contains an implicit geometrical assumption: The scene geometry is planar and coincides with the focal plane. Deviation of the scene geometry from the focal plane causes image warping. If depth information for each view is available, a specific geometrical warping can compensate the image distortion. Heidrich et al. [8] introduce a

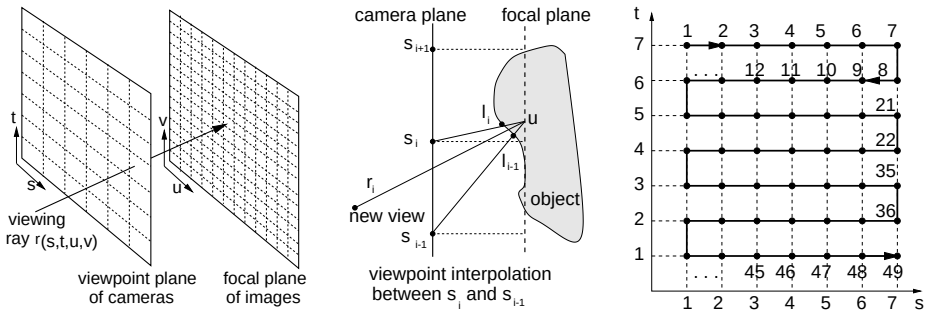

Fig. 1. Left: 4-D lightfield data structure with (s,t) viewpoint plane and (u,v) focal plane. Center: Rendering of novel views by interpolation of viewing ray r between the grid coordinates in the (s,u) slice. The radiance is interpolated from the object radiance at positions $l(s_i,u)$ and $l(s_{i-1},u)$. Image distortion occurs if the object surface deviates from the focal plane. Right: Tracking path for camera calibration along the (s,t)-viewing grid. Measurements are performed sequentially from image 1 to 49.

warping-based refinement from a depth-compensated lightfield to synthesize intermediate views. They construct a dense lightfield from a sparse set of ray-traced synthetic images. This approach allows interactive visualization of complex ray-traced scenes that is split into the initial off-line ray-tracing of few images and the online refinement for lightfield rendering. The problem is facilitated by the fact that calibration and depth estimation is obsolete since we deal with synthetic scenes. The ray-tracer delivers all necessary depth information as side product of the rendering. The discussion above reveals two major problems when acquiring lightfields from real image sequences:
– the need to directly obtain camera calibration from the image data, and
– the need to estimate local depth for view interpolation.

The original lumigraph approach [6] already tackles both problems. A calibration of the camera is obtained by incorporating a background with a known calibration pattern into the scene. The known specific markers on the background are used to obtain camera parameters and pose estimation [19]. It provides no means to calibrate the images from image data only. For depth integration the object geometry is approximated by constructing a visual hull from the object silhouettes. The hull approximates the global surface geometry but can not deal with local concavities. Furthermore, the silhouette approach is not feasible for general scenes and viewing conditions since a specific background is needed. This approach is therefore confined to laboratory conditions and does not provide a general solution for arbitrary scenes.

3 Camera calibration and geometric reconstruction

The problem of simultaneous camera calibration and depth estimation from image sequences (structure-from-motion, SFM) has been addressed for quite some time in the computer vision community. In the case of known intrinsic camera parameters, the camera pose as well as the scene structure can be estimated

from correspondences in the 2D image sequence up to an unknown scale factor. Longuet–Higgins [15] first demonstrated how to obtain structure and camera pose from eight point correspondences in one image pair. The uniqueness of this external calibration was proven in [18]. It exploits the basic relationship between image correspondences of a rigid scene, the Essential matrix E. The approach has been extended in several works, e.g. [10, 4] to an arbitrary number of point correspondences and views using non-linear optimization methods. Faugeras [5] and Hartley [7] later demonstrated that a projective reconstruction is possible from image matches alone even if the camera is totally uncalibrated.

A 3D scene reconstruction system using structure-from-motion was proposed by Beardsley et al. [1] who obtained projective calibration and sparse 3D structure by robustly tracking salient feature points throughout an image sequence. We have extended their method to obtain metric reconstructions (Euclidean reconstruction up to global scale) for fully uncalibrated sequences with methods of self-calibration [16]. For dense structure recovery a stereo matching technique was applied between image pairs of the sequence to obtain a dense depth map for each viewpoint. From this depth map a triangular surface wire-frame is constructed and texture mapping from the image is applied to obtain realistic surface models [11]. To summarize, we obtain a metric scene reconstruction in a 3-step approach:

1. Camera pose calibration is obtained by robust tracking of salient feature points over the image sequence,
2. local dense depth maps for all viewpoints are computed from correspondences between adjacent image pairs of the sequence,
3. a global 3-D surface mesh approximates the geometry, and surface texture is mapped onto it to enhance the visual appearance.

3.1 Combining Lightfield rendering and SFM

If we compare lightfield rendering and SFM, we see a considerable overlap. Both approaches require a good camera calibration and the estimation of local depth maps from the image data. For a geometric reconstruction we then need to combine all local depth estimates into a globally consistent surface model with a unique surface texture. This may be difficult to obtain for complex geometries and reflectivities. It would be better to compute depth maps only and to switch the geometry and surface texture depending on the current rendering viewpoint. And this is precisely what the lightfield approach can do once calibration and depth maps are given [8]. We therefore propose to combine the first two steps of our structure-from-motion approach with lightfield rendering.

The calibration is facilitated for the lightfield approach since we use densely spaced viewpoints where the adjacent images are rather similar. The camera viewpoints are tracked sequentially (along the 1-D viewing path that the camera takes). However, for a lightfield we are obtaining a 2-D viewing surface with the camera viewpoints as nodes of this grid in the s, t-plane. With a moving camera we can scan this viewing surface row by row in a sequential fashion (see

also Fig. 1, right). The camera poses are estimated by tracking salient image features throughout the sequence. Salient image feature points are matched using robust (RANSAC) techniques for that purpose. At first feature correspondences are found by extracting intensity corners in different images and matching them using a robust corner matcher [17]. In conjunction with the corner matching a restricted calibration of the setup is calculated. This allows to eliminate matches which are inconsistent with the calibration. The matching is started on the first two images of the sequence. The calibration of these views defines a metric coordinate system in which the projection matrices of the other views are retrieved one by one. A depth triangulation of the corresponding image matches will give a 3D estimate of salient scene points. In subsequent views we utilize this 3D estimate to predict correspondences and to verify them throughout the sequence. The estimated 3D feature points also define a coarse estimate of 3D scene structure. The intrinsic camera parameters were calibrated offline [19] and the approach of [16] has been modified to estimate the camera poses only. This allows robust metric reconstruction from any camera motion.

Once we have retrieved the metric calibration of the cameras we can use image correspondence techniques to estimate scene depth. For dense correspondence matching a disparity estimator based on the dynamic programming scheme of Cox *et al.* [2] is employed. It operates on rectified image pairs where the epipolar lines coincide with image scan lines. The rectification is easily obtained for each pair of adjacent viewpoints by projective mapping of the image planes onto standard parallel stereo geometry. The matcher searches at each pixel in one image for maximum normalized cross correlation in the other image by shifting a small measurement window (kernel size 5x5 or 7x7) along the corresponding scan line. The algorithm employs an extended neighborhood relationship and a pyramidal estimation scheme to reliably deal with very large disparity ranges [3]. It was further extended to multi-viewpoint depth analysis [11]. This allows to obtain locally consistent dense depth estimates for each viewpoint.

4 Experimental Results

In order to test the approach we used a calibrated robot arm for image acquisition. This allows us to obtain ground truth information for the camera pose. The intrinsic parameters were estimated off-line before the experiments.

The camera is mounted on the arm of a robot of type SCORBOT-ER VII. The position of its picker arm is known from the angles of the 5 axes and the dimensions of the arm. Optical calibration methods have to be applied to determine the relative position of the camera to the picker arm. This is done by the hand/eye–calibration method of [20]. The main problem is to determine the position along the optical axis of the camera. The repetition error of the robot is 0.2 mm and 0.03 degrees, respectively. Because of the limited size of the robot, we are restricted to scenes with maximum size of about 100 mm in diameter. For testing we used a scene with planar motion (compliant to the viewpoint plane) and a scene with spherical motion.

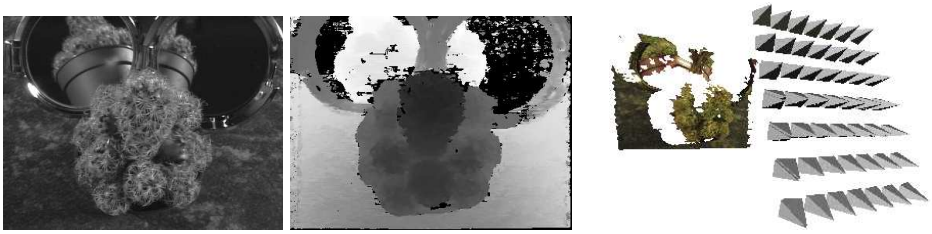

Fig. 2. Planar motion on a 7×7 viewpoint grid. Left: One of the input images. Middle: Corresponding dense depth map (color coded: dark=near,light=far, black=undefined). Right: 3D surface model and calibrated camera grid. The little pyramids symbolize the estimated camera poses.

Planar motion: To show the applicability for natural scenes, we chose a small cactus together with two mirrors. This scene has a non-trivial geometry with occlusions, spikes, reflections, etc., see fig. 2. In the planar case, we controlled the robot so that the center of projection moved within a planar 7×7 grid, the optical axis always intersecting one central point in the middle of the scene. The grid had spacing between the views of 13.3×16.6 mm. The orthogonal distance of the grid to the central point was 200 mm. This setup allows ground truth comparison of the camera calibration method. During calibration we tracked the camera positions sequentially row by row, moving like a snake from image 1 to 49 over all views (see fig. 1,right). We did not consider the connectivity between the rows which would additionally stabilize the tracking. The calibrated sequence allows to reconstruct dense depth maps for each viewpoint. From the depth map we obtain local geometry for image warping. Results of the 3D surface modeling are shown in fig. 2.

A quantitative evaluation of the camera calibration can be computed if we compare the length of the estimated displacements between adjacent camera positions (the baseline) with the baseline as given by the robot. Since the SFM algorithm generates only metric estimates (arbitrary scale for the baseline between the first two cameras) we scaled the estimates to the known robot baseline of 13.3 mm between the first two cameras. We could then measure all baselines between adjacent cameras.

The result is summarized in table 4. We have to distinguish between the statistics of row and column displacements. Since the camera moved sequentially row by row (see fig. 1,right), only the camera positions along the rows were estimated directly. The adjacency between columns was not exploited which causes an increased column baseline error. Still, the column statistics show a very good agreement with the expected value and no significant error accumulation was noticed. These figures document the stability and accuracy of the proposed calibration method. The overall performance of the calibration is within the range of the robot arm accuracy.

Spherical motion The proposed system can work with any camera motion and is not confined to planar viewing planes. To test this we performed a spherical robot motion. In the spherical case the robot sampled a 8×8 spherical grid

Table 1. Statistical distribution of the baseline length between camera positions.

Displacement Statistics	ground truth robot baseline[mm]	measured baselines mean [mm]	standard deviation [mm]	repeatability of robot [mm]
rows	13.30	12.65	0.478	0.2
columns	16.60	16.99	0.931	0.2

with a radius of 230 mm. The viewing positions enclosed a maximum angle of 45 degrees. The results of the camera calibration and geometric estimation are shown in fig. 3. The estimated camera positions are equally spaced on the viewing sphere and the geometry shows quite some detail.

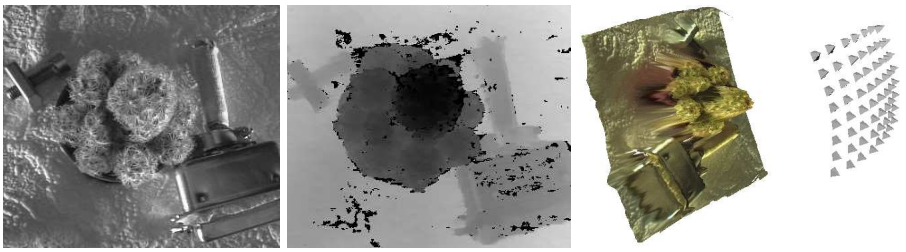

Fig. 3. Spherical motion. Left: One of the input images. Center: Dense depth map of scene. Right: 3D surface model and calibrated camera grid of spherical scene.

5 Conclusions and Further Work

We have employed a geometric structure-from-motion approach for lightfield calibration and local depth estimation which can be used to improve lightfield rendering. Image acquisition as well as rendering quality will profit from this integration. Most notably, we were able to generate lightfields from image sequences of freely moving cameras.

We are currently working to further integrate both approaches. The ongoing research has delivered additional results which we could not include here but would like to refer to. The inherent two-dimensional relationship between lightfield images has been exploited, resulting in a robust calibration of a 2D viewpoint mesh over all views [12] and improved depth reconstruction from the viewpoint mesh [13]. The image-based rendering approach has been adapted to incorporate local depth estimates and to render images from irregular viewpoint meshes of hand-held camera sequences [9].

Acknowledgments

We gratefully acknowledge financial support from the Belgian project IUAP 04/24 'IMechS', and the German Science Foundation DFG SFB-603,C2.

References

1. P. Beardsley, P. Torr and A. Zisserman: 3D Model Acquisition from Extended Image Sequences. In: B. Buxton, R. Cipolla(Eds.) Computer Vision - ECCV 96, Cambridge, UK., vol.2, pp.683-695. LNCS, Vol. 1064. Springer, 1996.
2. I. J. Cox, S. L. Hingorani, and S. B. Rao: A Maximum Likelihood Stereo Algorithm. Computer Vision and Image Understanding, Vol. 63, No. 3, May 1996.
3. L.Falkenhagen: Hierarchical Block-Based Disparity Estimation Considering Neighborhood Constraints. Intern. Workshop on SNHC and 3D Imaging, Rhodes, Greece, Sept. 1997.
4. O. Faugeras, S.Maybank: Motion from Point Matches: Multiplicity of solutions. IJCV 4(3), June 1990.
5. O. Faugeras: What can be seen in three dimensions with an uncalibrated stereo rig. Proc. ECCV'92, pp.563-578.
6. S. Gortler, R. Grzeszczuk, R. Szeliski, M. F. Cohen: The Lumigraph. Proceedings SIGGRAPH '96, pp. 43–54, ACM Press, New York, 1996.
7. R. Hartley: Estimation of relative camera positions for uncalibrated cameras. Proc. ECCV'92, pp.579-587.
8. W. Heidrich, H. Schirmacher, and H.-P. Seidel: A Warping-Based Refinement of Lumigraphs. Proc. WSCG '99, N.M. Thalman and V. Skala(eds.), Pilsen, 1999.
9. B. Heigl, R. Koch, M. Pollefeys, J. Denzler, L. Van Gool: Plenoptic Modeling and Rendering from Image Sequences taken by a Hand–Held Camera. Submitted to DAGM symposium, Bonn, Germany.
10. T.S. Huang, A.N. Netravali: 3D Motion estimation. in: H. Freeman (ed.), Machine vision for three-dimensional scenes, Academic Press, 1990.
11. R. Koch, M. Pollefeys, and L. Van Gool: Multi Viewpoint Stereo from Uncalibrated Video Sequences. Proc. ECCV'98, Freiburg, June 1998.
12. R. Koch, B. Heigl, M. Pollefeys, L. Van Gool, H. Niemann: Calibration of Handheld Camera Sequences for Plenoptic Modeling. Submitted to ICCV'99, Corfu.
13. R. Koch, M. Pollefeys, L. Van Gool: Robust Calibration and 3D Geometric Modeling from Large Collections of Uncalibrated Images. Submitted to DAGM Symposium 99, Bonn, Germany.
14. M. Levoy, P. Hanrahan: Lightfield Rendering. Proceedings SIGGRAPH '96, pp. 31–42, ACM Press, New York, 1996.
15. H.C. Longuet–Higgins: A computer algorithm for reconstructing a scene from two projections. Nature vol. 293 (81), pp. 133–135, 1981.
16. M. Pollefeys, R. Koch and L. Van Gool: Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters. Proc. ICCV'98, Bombay, India, Jan. 1998. Also accepted for publication in: International Journal on Computer Vision, Marr Price Special Issue, 1998.
17. P.H.S. Torr: Motion Segmentation and Outlier Detection. PhD thesis, University of Oxford, UK, 1995.
18. R. Y. Tsai, T. S. Huang: Uniqueness and Estimation of Three–Dimensional Motion Parameters of Rigid Objects with Curved Surfaces. PAMI 6(1), pp. 13–27, 1984.
19. R.Y.Tsai: A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology using off-the-shelf Cameras and Lenses. IEEE Journal Robotics and Automation RA-3,4 (Aug. 1987), 323-344.
20. R.Y. Tsai and R.K. Lenz: Real Time Versatile Robotics Hand/Eye Calibration Using 3D machine vision. Proceedings of ICRA 88, 554–561, IEEE Press, Philadelphia, April 1998.