Robust Calibration and 3D Geometric Modeling from Large Collections of Uncalibrated Images

Reinhard Koch[1], Marc Pollefeys[2], and Luc Van Gool[2]

[1] Multimedia Systems, Inst. of Computer Science, University of Kiel, Germany***
[2] Center for Processing of Speech and Images, K.U.Leuven, Belgium

Abstract. In this contribution we focus on calibration and 3D surface modeling from uncalibrated images. A large number of images from a scene is collected with a hand-held camera by simply waving the camera around the objects to be modeled. The images need not be taken in sequential order, thus either video streams or sets of still images may be processed. Since images are taken from all possible viewpoints and directions, we are effectively sampling the viewing sphere around the objects.

Viewpoint calibration is obtained with a structure-from-motion approach that tracks salient image points over multiple images. The calibration exploits the topology of the viewpoint distribution over the viewing sphere and builds a viewpoint mesh that connects all nearby viewpoints, resulting in a robust multi-image calibration. For each viewpoint a depth map is estimated that considers all corresponding image matches of nearby viewpoints. All depth maps are fused to generate a viewpoint-independent 3D surface representation based on a volumetric voting scheme. A voxel space is built into which the depth estimates from all the viewpoints are projected, together with their estimation uncertainty. Integration over all depth estimates determines a probability density distribution of the estimated scene surface. The approach was verified on long image sequences obtained with a hand-held video camera.

1 Introduction

In this contribution we discuss the evaluation of large collection of images for the purpose of 3D scene reconstruction. A large number of images from a scene is collected with a hand-held camera by simply waving the camera around the objects to be modeled. Since images are taking from all possible viewpoints and directions, we are effectively sampling the viewing sphere around the objects and generating a mesh of viewpoints. The collection of images can be exploited to reconstruct the scene, either with image-based rendering techniques [5] or through reconstruction of 3D surface geometry.

This work is embedded in the context of *uncalibrated Structure From Motion* (SFM) where camera calibration and scene geometry are recovered from images of the scene alone without the need for further scene or camera information.

*** Work performed while at K.U. Leuven

Faugeras and Hartley first demonstrated how to obtain uncalibrated projective reconstructions from image point matches alone [3,6]. Since then, researchers tried to find ways to upgrade these reconstructions to metric (i.e. Euclidean but unknown scale, see [4,15]). Beardsley et al. [1] proposed a scheme to obtain projective calibration and 3D structure by robustly tracking salient feature points throughout an image sequence. This sparse object representation outlines the object shape, but gives not sufficient surface detail for visual reconstruction. Highly realistic 3D surface models need a dense depth reconstruction and can not rely on few feature points alone.

In [11] the method of Beardsley e.a. was extended in two directions. On the one hand the projective reconstruction was updated to metric even for varying internal camera parameters, on the other hand a dense stereo matching technique [2] was applied between two selected images of the sequence to obtain a dense depth map for a single viewpoint. From this depth map a triangular surface wire-frame was constructed and texture mapping from one image was applied to obtain realistic surface models. In [8] the approach was further extended to multi viewpoint depth analysis. The approach can be summarized in 3 steps:

- Camera self-calibration and metric structure is obtained by robust tracking of salient feature points over the image sequence,
- dense correspondence maps are computed between adjacent image pairs of the sequence,
- all correspondence maps are linked together by multiple view point linking to fuse depth measurements over the sequence.

In [7,9,10] this approach was applied to the calibration of lightfield sequences from hand-held cameras. In this contribution we will extend our approach to the calibration of large image collections that sample the viewing sphere of the scene. Novel measures for determining the topological adjacency between viewpoints are developed in sect. 2. A volumetric surface reconstruction approach is introduced in sect. 3 that integrates all depth maps into a consistent 3D scene representation. Experiments on view calibration and geometric approximation conclude this contribution.

2 Calibration of a mesh of viewpoints

When very long image sequences have to be processed there is a risk of calibration failure due to several factors. For one, the calibration as described above is built sequentially by adding one view at a time. This may result in accumulation errors that introduce a bias to the calibration. Secondly, if a single image in the sequence is not matched, the complete calibration fails. Finally, sequential calibration does not exploit the specific image acquisition structure used in this approach to sample the viewing sphere. In this section we will develop a multi-viewpoint calibration algorithm that allows to actually weave the viewpoint sequence into a connected viewpoint mesh.

Image pair matching. The basic tool for viewpoint calibration is the two-view matcher. Image features have to be matched between the two images I_i, I_k of the viewpoints P_i, P_k. Here we rely on a robust computation of the Fundamental matrix F_{ik} with the RANSAC (RANdom SAmpling Consensus) method [14]. A minimum set of 7 features correspondences is picked from a large list of potential image matches to compute a specific F. For this particular F the support is computed from the other potential matches. This procedure is repeated randomly to obtain the most likely F_{ik} with best support in feature correspondence. From the F we can initialize a projective camera pair that defines a projective frame for reconstruction of the corresponding point pairs [1].

Once we have obtained the projection matrices we can triangulate the corresponding image features to obtain the corresponding 3D object features. The object points are determined such that their reprojection error in the images is minimized. In addition we compute the point uncertainty covariance to keep track of measurement uncertainties. The 3D object points serve as the *memory* for consistent camera tracking, and it is desirable to track the projection of the 3D points through as many images as possible. This process is repeated by adding new viewpoints and correspondences throughout the sequence. Finally constraints are applied to the cameras to obtain a metric reconstruction. A detailed account of this approach can be found in [12, 13].

Estimating the viewpoint topology. Since we are collecting a large amount of images from all possible viewpoints distributed over the viewing sphere, it is no longer reasonable to consider a sequential processing along the sequence frame index alone. Instead we would like to evaluate the image collection in order to robustly establish image relationships between all nearby images. We need to define a distance measure that allows to estimate the proximity of two viewpoints from image matches alone. We are interested in finding those camera viewpoints that are near to the current viewpoint and that support calibration. Obvious candidates for these are the preceding and following frames in a sequence, but normally those viewpoints are taken more or less on a linear path due to camera motion. This near-linear motion may lead to degeneracies and problems in the calibration. We are therefore also interested in additional viewpoints that are perpendicular to the current direction of the camera motion. If the camera sweeps back and forth over the viewpoint surface we will likely approach the current viewpoint in previous and future frames. Our goal is now to determine which of all viewpoints are nearest and most evenly distributed around our current view. The measurement tool we have at hand is the F-Matrix-computation from corresponding image points. For each potential neighbor we compute F w.r.t. the current image. To measure *proximity* and *direction* of the matched viewpoint w.r.t. the current one, we can exploit the image epipole as well as the distribution of the correspondence vectors.

Proximity: The distribution of the corresponding matches determines the distance between two viewpoints. Consider a non-planar scene and general motion

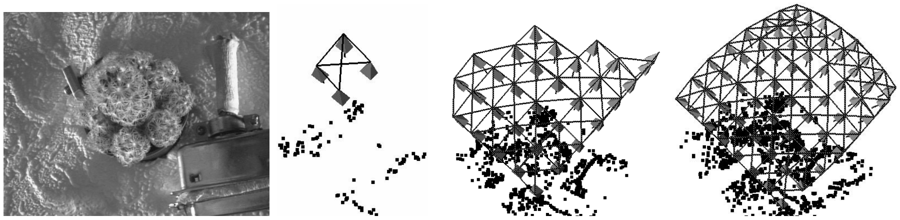

Fig. 1. Left: one image of the robot sequence. Left to right: Buildup of calibration after 4, 32, and 64 images. The camera viewpoints are indicated by pyramids that are connected by the viewpoint mesh. The black points in the background are tracked 3D feature points.

between both cameras. If both camera viewpoints coincide we can cancel out the camera orientation change between the views with a projective mapping (rectification) and the corresponding points will coincide since no depth parallax is involved. For a general position of the second camera viewpoint, the depth parallax will cause a residual correspondence error e_r after rectification that is proportional to the baseline distance between the viewpoints. We can approximate the projective rectification by a linear affine mapping that is estimated from the image correspondences. We therefore define the residual correspondence error e_r as proximity measure for nearby viewpoints.

Direction: The epipole determines the angular direction α_e of the neighboring camera position, since it represents the projection of the camera center into the image. Those viewpoints whose epipoles are most evenly distributed over all image quadrants should be selected for calibration.

Weaving the viewpoint mesh. With the distance measure at hand we can build a topological network of viewpoints from an unordered collection of images. This is necessary if one collects a large set of images of a scene with a still camera. Exploitation of the sequential frame index in a camera sequence however will give some speed advantages since we do not need to compute all possible F-Matrices. The strategy for sequential mesh building was described in detail in [10]. Here we extend that approach to non-sequential image collections.

We start with an arbitrary image of the sequence and compute α_e and e_r for subsequent images. If we choose the starting image as first image of the sequence, we can proceed along the frame index and find the nearest adjacent viewpoints in all directions. From this seed views we proceed recursively, building up the viewpoint mesh topology over all views. The method is visualized in fig. 1 for a sequence taken with a robot arm. The camera is mounted on the arm of a robot of type SCORBOT-ER VII. The robot sampled a 8×8 spherical viewing grid with a radius of 230 mm. The viewing positions enclosed a maximum angle of 45 degrees which gives an extension of the spherical viewpoint surface patch of $180 \times 180\,\mathrm{mm}^2$. The scene consists of a cactus and some metallic parts on a piece of rough white wallpaper. One of the original images is shown in fig. 1 together with the computed viewpoint mesh after tracking of 4, 32, and 64 images (from left to right). The mesh buildup is indicated by the estimated camera viewpoints (pyramids) and their topological relation (mesh connecting the cameras). Each connection indicates that the fundamental matrix between the image pair has been computed. The mesh builds along the shortest camera distances very much like a wave propagating over the viewpoint surface.

3 3D scene reconstruction

Once we have retrieved the metric calibration of the cameras we can use image correspondence techniques to estimate scene depth. We rely on stereo matching techniques that were developed for dense and reliable matching between adjacent views. The small baseline paradigm suffices here since we use a rather dense sampling of viewpoints.

3.1 Dense depth estimation by correspondence matching

For dense correspondence matching an area-based disparity estimator is employed. 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 7x7) along the corresponding scan line. Dynamic programming is used to evaluate extended image neighborhood relationships and a pyramidal estimation scheme allows to reliably deal with very large disparity ranges [2]. The geometry of the viewpoint mesh is especially suited for further improvement with a multi viewpoint refinement [8]. For each viewpoint a number of adjacent viewpoints exist that allow correspondence matching. Since the different views are rather similar we will observe every object point in many nearby images. This redundancy can also be exploited to verify the depth estimation for each object point, and to refine the depth values to high accuracy.

3.2 View-independent object modeling from multiple depth maps

The depth maps as obtained so far represent a robust and accurate estimate of local scene geometry. Care has been taken to eliminate measurement outliers from the depth map. Unfortunately a depth map is viewpoint-dependent and unable to represent occluded object parts. We have however a large collection of calibrated and registered depth maps at hand which can be converted into a consistent and viewpoint-independent 3D scene model.

Building voxel walls. A full 3D representation is possible if we employ a 3D data structure like the voxel space. For model building we define a 3D voxel volume that bounds the scene geometry. Since the depth maps are registered by the calibration, we can simply project them into the volume. Each depth estimate defines a 3D surface point with an associated depth uncertainty covariance that

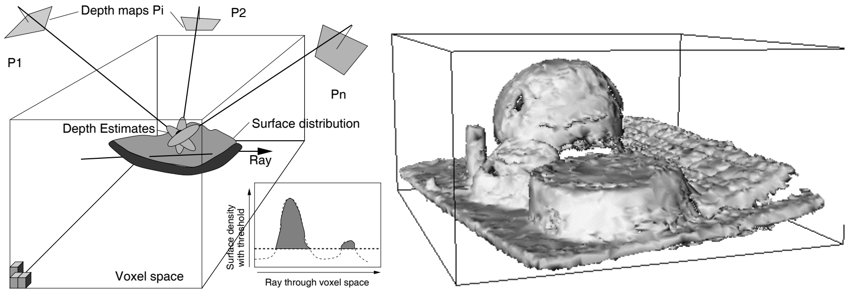

Fig. 2. Left: Voxel space with cameras projecting surface estimates into the voxel space. The surface is defined as the accumulated probability density distribution value above a certain threshold. Right: results of surface segmentation for the office scene (see fig 3). The surface was reconstructed from 151 depth maps projected into a 128^3 voxel space.

determines the probability distribution of the point. A voxel will then represent the probability of the estimated 3D surface point. The integration of all available depth estimates in voxel space builds a 3D probability density volume with the maxima being the most likely 3D surface points. Fig. 2 demonstrates the mapping of depth estimates into the voxel space.

The defined voxel resolution quantizes the surface distribution into nearest-neighbor approximations. If the voxel quantization is coarser than the estimation uncertainty, then the density projection approach is reduced to simply incrementing individual voxel values. We can think of this technique as building a wall by setting all individual stones (voxels) of the wall. Each surface voxel will be hit more than once because it is exposed to multiple views. The probability of a surface voxel is therefore high as compared to interior and exterior points. Thus we obtain a robust hough-like integration scheme for surface point candidates where most of the hits are concentrated on the 3D surface. Outliers will hit wrong voxels but they are easily discarded by thresholding the voxel distribution.

Volume-Boundary representations. The volume density distribution can be converted to a surface representation for further processing. The maxima of the distribution correspond to the most likely surface points. We can classify surface voxels by simply thresholding the distribution. All voxel values above a certain threshold are considered as surface points. A volume boundary conversion like marching cubes will then define a closed surface around the real surface points. The true surface is guaranteed to be inside the enclosed volume. For voxel quantizations where the depth uncertainty is of the order of the voxel size we obtain very reliable surfaces reconstructions. The surface extraction is sketched in Fig. 2 (left). It shows a ray through the volume with associated density distribution. The surface is selected with a global threshold that determines the minimum number of hits for a voxel in order to be classified as surface. The figure 2 (right) displays a volumetric surface reconstruction from a real hand-held sequence, see the next section.

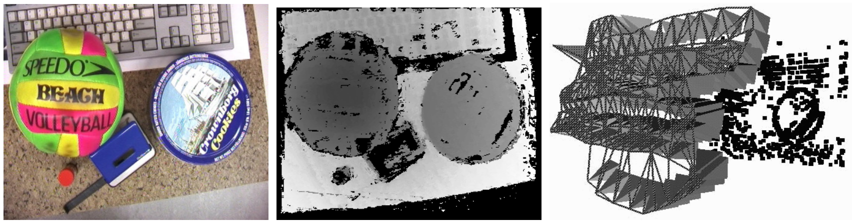

Fig. 3. (Image (left) and depth map (center) from hand-held office sequence. The depth map was obtained from 7 adjacent views. Right: Viewpoint mesh (in grey) with cameras as pyramids and tracked 3D points in the background (black).

4 Experiment: Hand-held office sequence

We tested our approach with an uncalibrated hand-held sequence. A digital consumer video camera (Sony DCR-TRV900 with progressive scan) was swept freely over a cluttered scene on a desk, covering a viewing surface of about 1 m^2. The resulting video stream was then digitized on an SGI O2 by simply grabbing 187 frames at more or less constant intervals. No care was taken to manually stabilize the camera sweep. Fig. 3(left) displays an images of the sequence and the corresponding depth map (middle). The tracked camera viewpoints and the viewpoint mesh topology is shown to the right.

Volumetric 3D representations. For the 3-D reconstruction we used 151 depth maps. Each map was already fused from adjacent images to reduce outliers and to improve the estimates. Fig. 4 shows various results of the reconstruction that demonstrate the high fidelity of our approach. Please note the very high detail resolution at the keyboard, the ball and the red glue stick. The stick was modeled from all directions which is impossible with depth map modeling alone. The surface can be textured as well to improve the visual result. A side view of the reconstruction is found in fig 2.

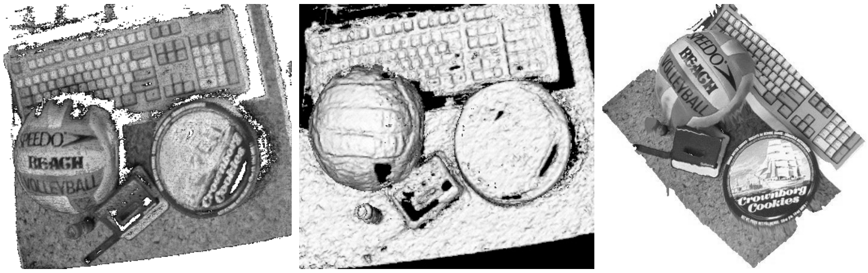

Fig. 4. Reconstruction results. Left: a volumetric reconstruction with a volume ray-caster where opacity and texture of each surface voxel was set to the mean color projected from all viewpoints. Center: a high resolution surface reconstruction obtained from a volume with a resolution of 256x256X128 voxel. Right: surface reconstruction with texture mapping.

5 Conclusions

We have described a system for the automatic calibration of large collections of images obtained with a hand-held video camera. The calibration exploits the proximity of viewpoints by building a viewpoint mesh that spans the viewing sphere around a scene. Once calibrated, the viewpoint mesh can be used for image-based rendering or 3D geometric modeling of the scene. We have further described a novel viewpoint-independent modeling approach that builds a 3D maximum likelihood surface estimate from the projection of depth estimates into the scene volume. It allows robust and highly accurate 3D scene modeling from uncalibrated hand-held image sequences.

References

1. P. Beardsley, P. Torr and A. Zisserman: 3D Model Acquisition from Extended Image Sequences. *ECCV 96*, LNCS 1064, vol.2, pp.683-695.Springer 1996.
2. L.Falkenhagen: Hierarchical Block-Based Disparity Estimation Considering Neighborhood Constraints. Intern. Workshop on SNHC and 3D Imaging, Rhodes, Greece, Sept. 1997.
3. O. Faugeras: What can be seen in three dimensions with an uncalibrated stereo rig. *Proc. ECCV'92*, pp.563-578.
4. O. Faugeras, Q.-T. Luong and S. Maybank: Camera self-calibration - Theory and experiments. *Proc. ECCV'92*, pp.321-334.
5. S. Gortler, R. Grzeszczuk, R. Szeliski, M. F. Cohen: The Lumigraph. Proceedings SIGGRAPH '96, pp 43–54, ACM Press, New York, 1996.
6. R. Hartley: Estimation of relative camera positions for uncalibrated cameras. *ECCV'92*, pp.579-587.
7. B. Heigl, R. Koch, M. Pollefeys, J. Denzler: Plenoptic Modeling and Rendering from Image Sequences taken by a Hand-Held Camera. Proceedings DAGM'99, Bonn, Sept. 1999.
8. R. Koch, M. Pollefeys, and L. Van Gool: Multi Viewpoint Stereo from Uncalibrated Video Sequences. *Proc. ECCV'98*, Freiburg, June 1998.
9. R. Koch, B. Heigl, M. Pollefeys, L. Van Gool, H. Niemann: A Geometric Approach to Lightfield Calibration. Proceedings CAIP'99, Ljubljana, Slovenia, Sept. 1999.
10. R. Koch, M. Pollefeys, B. Heigl, L. Van Gool, H. Niemann: Calibration of Handheld Camera Sequences for Plenoptic Modeling. Proc. of ICCV'99, Korfu, Greece, Sept. 1999.
11. 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.
12. M. Pollefeys, R. Koch, M. Vergauwen and L. Van Gool: Metric 3D Surface Reconstruction from Uncalibrated Image Sequences. In: 3D Structure from Multiple Images of Large Scale Environments. LNCS Series Vol. 1506, pp. 139-154. Springer-Verlag, 1998.
13. M. Pollefeys: Self-Calibration and Metric 3D Reconstruction from Uncalibrated Image Sequences. Ph.D. Thesis, K.U.Leuven, May 1999.
14. P.H.S. Torr: Motion Segmentation and Outlier Detection. PhD thesis, University of Oxford, UK, 1995.
15. B. Triggs: The Absolute Quadric. *Proc. CVPR'97.*