REALISTIC 3-D SCENE MODELING FROM UNCALIBRATED IMAGE SEQUENCES

Reinhard Koch[*]

Multimedia Information Processing
Institute of Computer Science
University of Kiel, Germany
rk@informatik.uni-kiel.de

Marc Pollefeys, Luc Van Gool

Center for Processing of Speech and Images
ESAT-PSI
Katholieke Universiteit Leuven, Belgium
{firstname.lastname}@esat.kuleuven.ac.be

ABSTRACT

This contribution addresses the problem of obtaining photo-realistic 3D models of a scene from images alone with a structure-from-motion approach. The 3D scene is observed from multiple viewpoints by freely moving a camera around the object. No restrictions on camera movement and internal camera parameters like zoom are imposed, as the camera pose and intrinsic parameters are calibrated from the sequence. The only restrictions on the scene content are the rigidity of the scene objects and opaque, piecewise smooth object surfaces. The approach operates independently of object scale and requires only a single low-cost consumer photo or video camera.

The modeling system described here uses a three-step approach. First, the camera pose and intrinsic parameters are calibrated on-line by tracking salient feature points between the different views. Next, consecutive images of the sequence are treated as stereoscopic image pairs and dense correspondence maps are computed by area matching. Finally, dense and accurate depth maps are computed by linking together all correspondences over the viewpoints. The depth maps are converted to triangular surfaces meshes that are texture mapped for photo-realistic appearance. The resulting surface models are stored in VRML-format for easy exchange and visualization.

The feasibility of the approach has been tested extensively and will be illustrated on several real scenes. In particular we will demonstrate the generation of realistic 3D models for a virtual exhibition of the archaeological excavation site in Sagalassos, Turkey.

Keywords: Structure from Motion, Camera calibration, 3-D Scene reconstruction, 3-D Modeling.

1. INTRODUCTION

The use of three-dimensional surface models for the purpose of visualization is gaining importance. Highly realistic 3-D models are readily used to visualize and simulate events, like in flight simulators, in the games and film industry or for product presentations. The range of applications span from architecture visualization over virtual television studios, virtual presence for video communications to general "virtual reality" applications.

A limitation to the widespread use of these techniques is currently the high cost of such 3-D models since they have to be produced manually. Especially if existing objects are to be reconstructed the measurement process for obtaining the correct geometric and photometric data is tedious and time consuming. Traditional solutions include the use of stereo rigs, laser range scanners and other 3-D digitizing devices. These devices are often very expensive, require careful handling and complex calibration procedures and are designed for a restricted depth range only.

To overcome the above mentioned problems we propose an image-based approach to 3-D scene modeling. The scene which has to be reconstructed is recorded from different viewpoints by a video camera. The relative position and orientation of the camera and its calibration parameters will automatically be retrieved from the image data by the algorithms. Hence, there is no need for measurements in the scene or calibration procedures whatsoever. There is also no restriction on range, it is just as easy to model a small object, as to model a complete landscape. The proposed method thus offers a previously unknown flexibility in 3-D model acquisition. In addition, any photographic recording device - e.g. cam-corder, digital camera, or even standard photographic film camera - is sufficient for scene acquisition. Hence, increased flexibility is accompanied by a decrease in cost.

In this contribution we will discuss the complete and automatic modeling system that is capable to reconstruct the scene from uncalibrated image sequences. We will then discuss applications to demonstrate the possible use of such a reconstruction system.

[*]Work performed while at the K.U. Leuven.

2. 3-D MODELING FROM VIDEO

2.1. State of the art

The proposed method is placed in the framework of uncalibrated scene reconstruction that is a recent research topic. In the uncalibrated case all parameters, camera pose and intrinsic calibration as well as the 3-D scene structure have to be estimated from the 2D image sequence alone. Faugeras and Hartley first demonstrated how to obtain uncalibrated projective reconstructions from image sequences alone [3, 4]. Since then, researchers tried to find ways to upgrade these reconstructions to metric (i.e. Euclidean but unknown scale). Newest results report full self-calibration methods even for varying intrinsic parameters like focal length, which allows the unrestricted use of the camera, for example zooming [8].

To employ these self-calibration methods for sequence analysis they must be embedded in a complete scene reconstruction system. Beardsley et al. [1] proposed a scheme to obtain projective calibration and 3-D 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 3-D surface models need the dense depth estimation and can not rely on few feature points alone.

In [8] 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 stereo matching 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 [5] the approach was further extended to multi-viewpoint sequence analysis.

2.2. System Overview

Robust camera calibration and accurate depth estimation are the key problems to be solved. In our system we use a 3-step approach that is visualized in fig. 1 with the example of modeling a building facade:

- Camera self-calibration is obtained by robust tracking of salient feature points over the image sequence,
- dense depth maps are computed between adjacent image pairs,
- depth maps are linked together by multiple view point linking to fuse depth measurements from all images into a consistent model. The model is stored as a textured 3-D surface mesh.

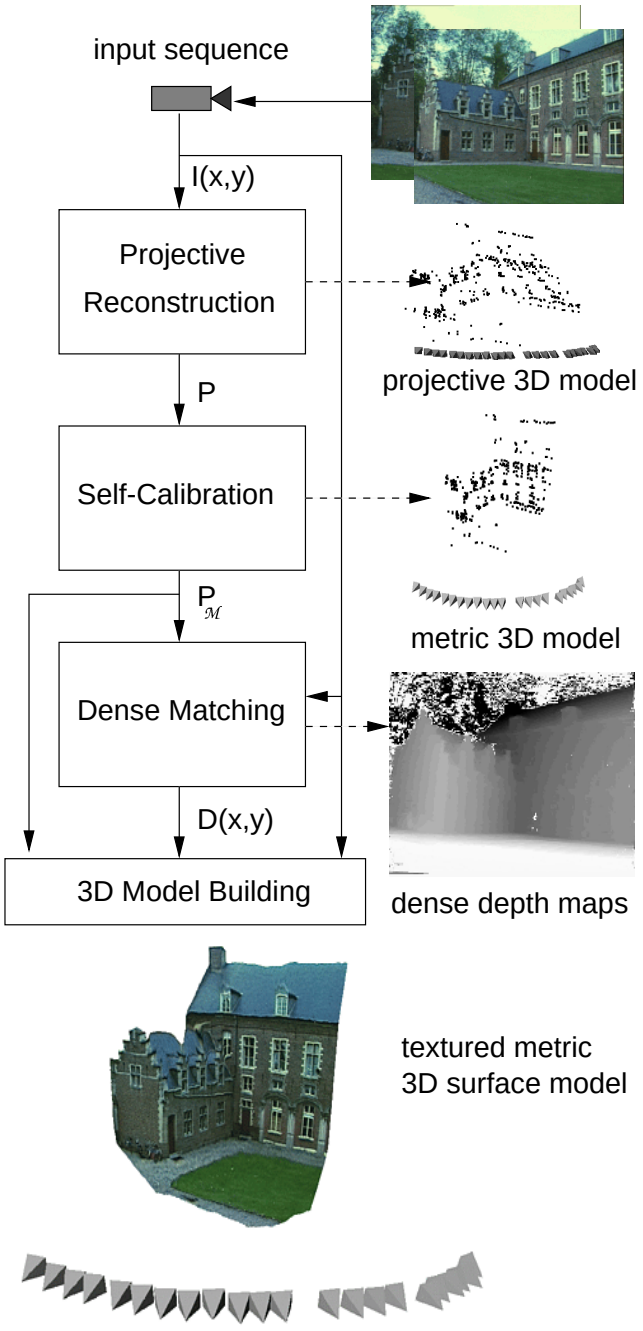

Figure 1: System overview: from the image sequence ($I(x,y)$) the projective reconstruction is computed; the projection matrices P are then passed on to the self-calibration module which delivers a metric calibration $P_{\mathcal{M}}$; the next module uses these to compute dense depth maps $D(x,y)$; all these results are assembled in the last module to yield a textured 3-D surface model. The little pyramids in fromt of the building symbolize the different camera positions.

2.3. Camera Calibration

Camera calibration[1] is obtained by tracking salient image features throughout the sequence. At first feature correspondences are found by extracting intensity corners in different images and matching them using a corner matcher with robust statistics (RANSAC) [10]. In conjunction with the matching of the corners a restricted calibration of the setup is calculated (i.e. only determined up to an arbitrary projective transformation). This allows to eliminate matches which are inconsistent with the calibration, by verifying the epipolar constraint. Using this constraint, more matches can easily be found and used to refine the calibration.

Reconstruction and calibration is initialized from the first two images of the sequence. The calibration of these views defines a projective framework in which the projection matrices of the other views are retrieved one by one. Salient points are tracked and compared with all adjacent views to generate a tight network of correspondences between all possible views. Thus we can retrieve the camera viewpoints with high accuracy and robustness from image matches only [6]. The image correspondences are then triangulated to form a sparse projective 3D reconstruction of salient points.

The projective reconstruction is determined only up to a projective transformation and forms a skewed representation of the metric world: orthogonality and parallelism are not preserved, part of the scene can be warped to infinity, etc (see projective reconstruction result in fig 1). To obtain a metric calibration, constraints on the internal camera parameters (e.g. absence of skew, known aspect ratio, ...) can be imposed for self-calibration. A detailed account of the methods employed in our system can be found in [8, 9].

2.4. Depth Estimation

Only a few salient scene points are reconstructed from feature tracking. Obtaining a dense reconstruction could be achieved by interpolation, but in practice this does not yield satisfactory results. Often some important features are missed during the corner matching and will not appear in the reconstruction. These problems can be avoided by using stereo matching algorithms which estimate correspondences for almost every point in the images. Since we have computed the calibration between successive image pairs we can exploit the epipolar constraint that restricts the correspondence search to a 1-D search range. In addition to the epipolar geometry, other constraints like preserving the order of neighboring pixels, bidirectional uniqueness of the match, and detection of occlusions can be included. These constraints are used for stereo matching of images pairs to guide the correspondence towards the most probable epipolar match using a dynamic programming scheme [2].

The pairwise disparity estimation allows to compute image correspondences between adjacent image pairs, and independent depth estimates for each camera viewpoint. An optimal joint estimate is achieved by fusing all independent estimates into a common 3-D model. The approach utilizes a flexible multi-viewpoint scheme by combining the advantages of small baseline and wide baseline stereo. Adjacent image pairs with small baselines are linked over the sequence by including more and more distant viewpoints to refine the estimate [5].

2.5. 3-D Modeling

Each depth map covers the scene structure that can be seen from this particular viewpoint only. To handle occlusions and to model a complete scene the different depth maps must be integrated in 3D-space. The fusion is obtained by building an intermediate 3-D volume into which all depth maps are projected while considering the estimation uncertainty of each depth estimate. The volume represents the probability density of the estimated 3-D scene surface [7].

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. Fig. 2 demonstrates the mapping of depth estimates into the voxel space.

Thresholding the regions of highest probability and con-

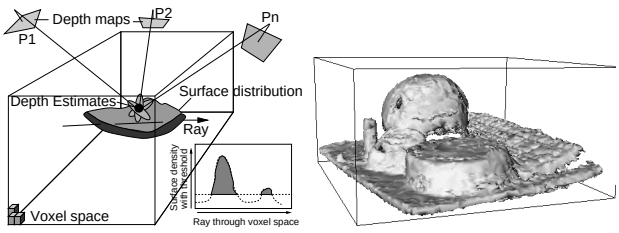

Figure 2: Left: cameras projecting surface estimates into the voxel space. The surface is defined as the accumulated probability density distribution value above a certain threshold. Right: surface reconstruction of the *office* scene, modeled from 187 views (see section 3.1).

[1] By *calibration* we mean the actual internal calibration of the camera as well as the relative position and orientation of the camera for the different views with respect to an arbitrary coordinate system.

Figure 3: Top: Image (left) and camera calibration (right) from hand-held office sequence. The little pyramids indicate the positions of the camera that was swept freely over the scene. Bottom left: Top view of surface geometry. Please note the detailed geometry of the keyboard, the ball and the glue stick. The stick was modeled from all sides which is impossible from a single viewpoint. Bottom right: Surface texturing of the model to improve the visual result.

verting those regions to a triangular wire-frame surface representation yields the 3-D scene geometry. The surface mesh is finally overlayed with a texture map taken from the real images to cover fine and un-modeled surface details. This results in a very realistic visual reconstruction quality.

3. APPLICATIONS

The proposed system was tested on a large variety of scenes with different cameras of varying quality (35 mm photo camera on Photo-CD, digital still camera, cam-corder) and was found to work even under difficult acquisition circumstances. We discuss two different examples.

3.1. 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(top left) displays an image of the sequence and the camera tracking (top right) with the cameras as pyramids and the tracked 3D points in the background (black). For the 3-D reconstruction we fused the

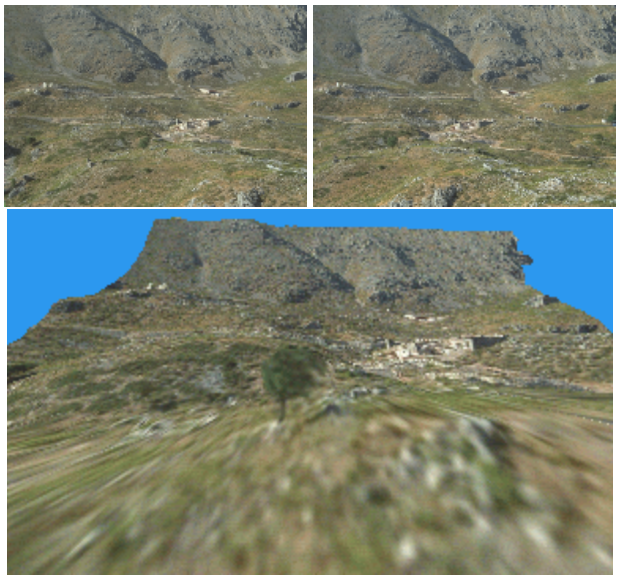

Figure 4: Images 2 and 9 of the site sequence (top) and overview model of the complete site (bottom).

depth maps from all viepoints. Fig. 3 (bottom) shows the resulting model from a top view, and a side view of the surface geometry can be seen in fig. 2 (right).

3.2. Sagalassos archaeological excavation site

Extensive field trials were carried out at the archaeological excavation site of Sagalassos in Turkey. This is a challenging task since the archaeologists want to reconstruct even small surface details and irregular structures, and the terrain is difficult. The goal of this field test was to prove the feasibility of our approach for a variety of scenes and to model objects for a virtual exhibition that can be presented over the Internet.

3.2.1. Sagalassos Site

The *Site* sequence in figure 4 is a good example of a large scale modeling using off-the-shelf equipment. Nine images of the complete excavation site (extension a few km^2) were taken with a conventional photographic camera while walking along the valley rim. The film was then digitized on Photo-CD. The *Site* reconstruction in figure 4 (bottom) gives a good overview of the valley relief. Some of the dominant objects like the Roman Bath and the Market place, as well as landmarks like big trees or stones are already modeled at this coarse scale but without any detail.

3.2.2. Detail models of the Roman bath

This sequence is a typical example of a detailed model. It consists of one part of the Roman bath that was modeled

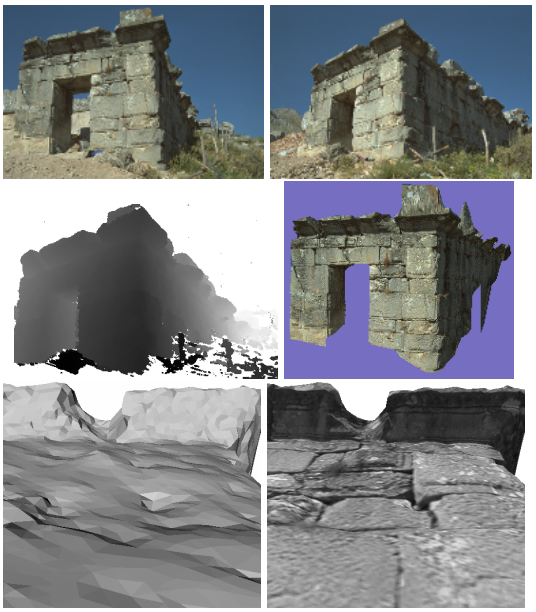

Figure 5: Top: 2 images of Roman Bath sequence. Middle: estimated depth map and reconstructed surface. Bottom: Detail of model without and with surface texture.

with high resolution. Figure 5 shows 2 of the six original images, the fused depth map and 3D-reconstructions. The depth reconstruction is very dense, and the relative depth error was estimated to 0.8%. Figure 5 (bottom) reveals the high reconstruction quality which gives a realistic impression of the object. The close-up view confirms that each stone is modeled, including relief and small indentations. The indentations belong to erosion gaps between the stones.

3.2.3. Augmenting the site model

With 3-D reconstructions at hand we can easily merge the reconstructed real site models with hypothesized buildings that were constructed from archaeological findings. This technique allows to visualize the site as it once might have been. In the case of Sagalassos some building hypothesis were translated to CAD models and integrated with our reconstructions. The site can then be visited in a *virtual tourist* application[2] where a virtual tour guide explains the site to a visitor, see fig. 6.

4. CONCLUSIONS

An automatic 3D scene reconstruction system was described that is capable to model textured surfaces from images of a freely moving, uncalibrated camera. It extracts metric surface models without prior knowledge about the scene or the camera other than assuming rigidity of the objects. The approach was tested with different off-the-shelf camera types on a variety of real scenes of varying scale and complexity.

[2] Reconstruction results can be viewed at http://www.esat.kuleuven.ac.be/psi/visics/demos.html

Figure 6: Sagalassos virtual tourism application combining video-based reconstructions, CAD models, and a virtual guide.

5. REFERENCES

[1] P. Beardsley, P. Torr and A. Zisserman: 3D Model Acquisition from Extended Image Sequences. ECCV 96, Springer LNCS 1064, Cambridge UK, 1996.

[2] L.Falkenhagen: Hierarchical Block-Based Disparity Estimation Considering Neighborhood Constraints. Int. WS SNHC and 3D Imaging, Rhodes, Greece, 1997.

[3] O. Faugeras: What can be seen in three dimensions with an uncalibrated stereo rig? ECCV 92,Springer LNCS 588.

[4] R. Hartley: Estimation of relative camera positions for uncalibrated cameras. ECCV 92,Springer LNCS 588, 1992.

[5] R. Koch, M. Pollefeys, and L. Van Gool: Multi-Viewpoint Stereo from Uncalibrated Video Sequences. ECCV 98, ,Springer LNCS 1406, 1998.

[6] R. Koch, M. Pollefeys, B. Heigl, L. Van Gool, H. Niemann: Calibration of Hand-held Camera Sequences for Plenoptic Modeling. Proc. ICCV 99, Korfu, Greece, 1999.

[7] R. Koch, M. Pollefeys, and L. Van Gool: Robust Calibration and 3D Geometric Modeling from Large Collections of Uncalibrated Images. Proc. DAGM 99, Bonn, Germany, 1999.

[8] M. Pollefeys, R. Koch and L. Van Gool: Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters. ICCV 98, Bombay, India, 1998.

[9] M. Pollefeys: Self-Calibration and Metric 3D Reconstruction from Uncalibrated Image Sequences. PhD Thesis K.U. Leuven, Mai 1999.

[10] P.H.S. Torr: Motion Segmentation and Outlier Detection. PhD thesis, University of Oxford, UK, 1995.