Automatic Generation of 3D Models from Photographs

Marc Pollefeys, Reinhard Koch, Maarten Vergauwen, Luc Van Gool
ESAT-PSI, K.U.Leuven, Kard. Mercierlaan 94,
B-3001 Leuven, Belgium
firstname.lastname@esat.kuleuven.ac.be

Abstract. In this paper a system is presented which automatically extracts a 3D surface model from a sequence of photographs of a static scene. The system can deal with unknown camerasettings. In addition the parameters of this camera are allowed to change during acquisition (e.g. by zooming or focussing). No prior knowledge about the scene is necessary to build the 3D models. This system therefore offers a high degree of flexibility. The system is based on state-of-the-art algorithms recently developed in computer vision. The 3D modelling task is decomposed into a number of succesive steps. Gradually more knowledge of the scene and the camera setup is retrieved.

1 Introduction

In the last few years the interest in 3D models has dramatically increased. More and more applications are using computer generated models. The main difficulty lies with the model acquisition. Although more tools are at hand to ease the generation of models, it is still a time consuming and expensive process. In many cases models of existing scenes or objects are desired. Traditional solutions include the use of stereo rigs, laser range scanners and other 3D digitizing devices. These devices are often very expensive, require careful handling and complex calibration procedures and are designed for a restricted depth range only.

In this work an alternative approach is proposed which avoids most of the problems mentioned above. The scene which has to be modeled is photographed from different viewpoints. The relative position and orientation of the camera and its calibration parameters will automatically be retrieved from image data. 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 (using a macro lens), as to model a complete building or even a whole landscape. The proposed method thus offers a previously unknown flexibility in 3D model acquisition. In addition, no more than a photo camera is needed for scene acquisition. Hence, increased flexibility is accompanied by a decrease in cost.

This flexibility opens the way to new applications. Scenes can be reconstructed from a sequence of photographs. Models can be generated from old film footage (e.g. from monuments destroyed during the war). It will become possible to generate realistic 3D models of complete sites (e.g. archeological sites, see [6]). Besides this, 3D modeling of objects (e.g. for tele-shopping applications or virtual exhibitions) is eased a lot.

2 Model Acquisition

Two things are needed to build a 3D model from an image sequence: (1) the calibration of the camera setup[1] and (2) the correspondences between the images. Starting from an image sequence acquired by an uncalibrated video camera, both these prerequisites are unknown and therefore have to be retrieved from image data. At least a few correspondences are needed to retrieve the calibration of the camera setup, but on the other hand this calibration facilitates the search for correspondences a lot.

In figure 1 an overview of the systems is given. It consists of independent modules which pass on the necessary information to the next modules. The first module computes the projective calibration of the sequence together with a sparse reconstruction. In the next module the metric calibration is computed from the projective camera matrices through self-calibration. Then dense correspondence maps are estimated. Finally all results are integrated in a textured 3D surface reconstruction of the scene under consideration. A more detailed description of this system can be found in [7].

2.1 Retrieving the Projective Framework

The first correspondences are found by extracting intensity corners in different images and matching them using a robust tracking algorithm. In conjunction with the matching of the corners the projective calibration of the setup is calculated. This allows to eliminate matches which are inconsistent with the calibration. Using the projective calibration more matches can easily be found and used to refine this calibration.

At first corresponding corners in two images are matched. This defines a projective framework in which the projection matrices of the other views are retrieved one by one. Our approach follows the procedure proposed by Beardsley *et al* [1]. We therefore obtain projection matrices (3×4) of the following form:

$$\mathbf{P}_1 = [\mathbf{I}|0] \text{ and } \mathbf{P}_k = [\mathbf{H}_{1k}|e_{1k}] \tag{1}$$

with $\mathbf{H}_{1k}$ the homography for some reference plane from view 1 to view k and e_{1k} the corresponding epipole.

2.2 Retrieving the Metric Framework

Such a projective calibration is certainly not satisfactory for the purpose of 3D modeling. A reconstruction obtained up to a projective transformation can differ very much from the original scene according to human perception: orthogonality and parallellism are in general not preserved, part of the scene can be warped to infinity, etc. To obtain a more complete calibration, constraints can be obtained by imposing some restrictions on the internal camera parameters (e.g. square pixels). By exploiting these constraints, the projective reconstruction can be upgraded to metric (Euclidean up to scale) [8, 9].

In a metric frame $\mathbf{P}$ can be expressed as follows:

$$\mathbf{P} = \mathbf{K}\,[\mathbf{R}|\text{-}\mathbf{R}\mathbf{t}] \text{ with } \mathbf{K} = \begin{bmatrix} f_x & s & u \\ & f_y & v \\ & & 1 \end{bmatrix} . \tag{2}$$

[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.

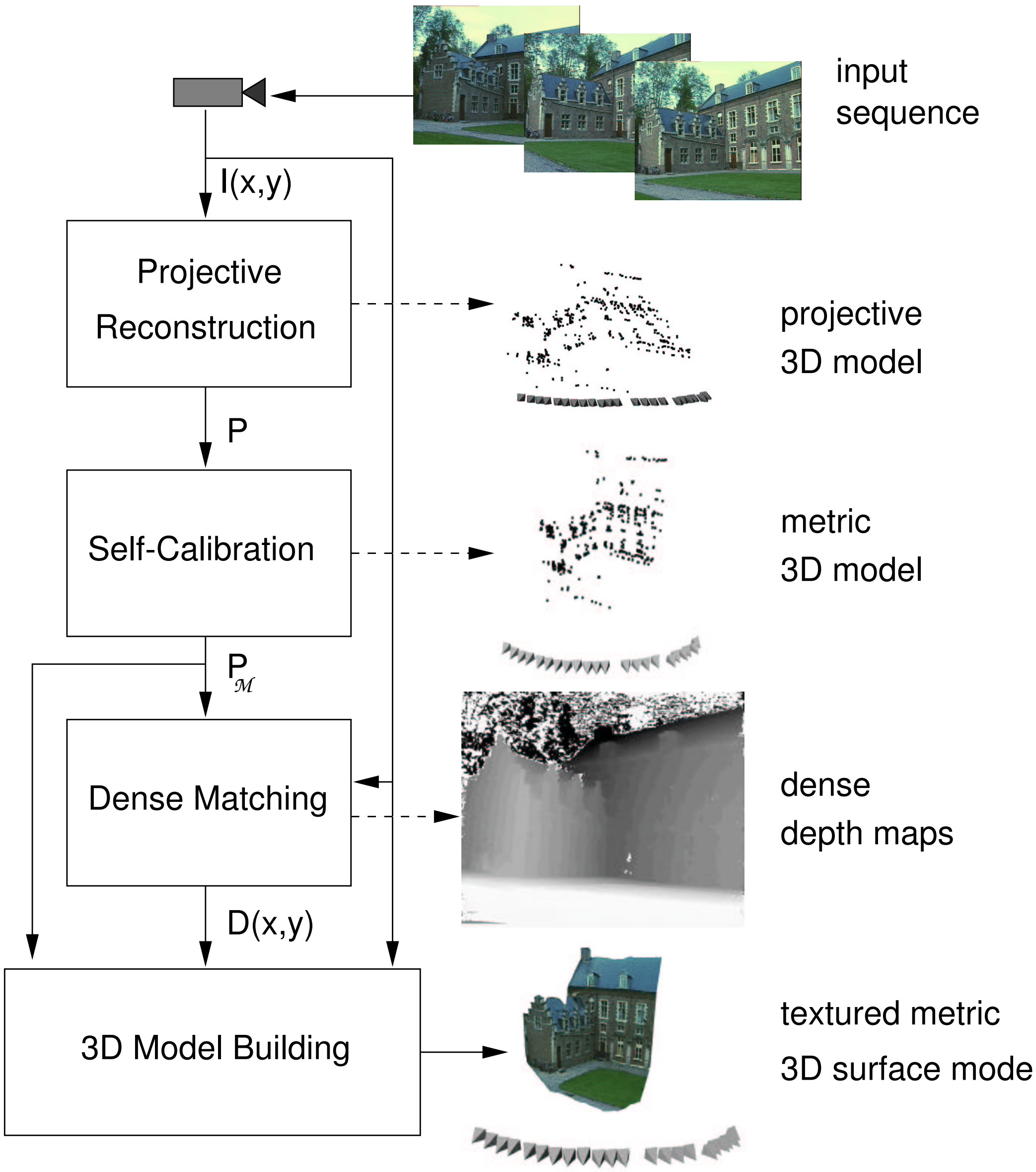

Figure 1: Overview of the system: from the image sequence $(I(x, y))$ the projective reconstruction is computed; the projection matrices P are then passed on to the self-calibration module which delivers a metric calibration $P_{\mathcal{M}}$; the next module uses these to compute dense depth maps $D(x, y)$; all these results are assembled in the last module to yield a textured 3D surface model. On the right side the results of the different modules are shown: the preliminary reconstructions (both projective and metric) are represented by point clouds, the cameras are represented by little pyramids, the results of the dense matching are accumulated in dense depth maps (light means close and dark means far).

Here $(\mathbf{R}, \mathbf{t})$ denotes a rigid transformation (i.e. $\mathbf{R}$ is a rotation matrix and $\mathbf{t}$ is a translation vector), while the upper triangular calibration matrix $\mathbf{K}$ encodes the intrinsic parameters of the camera (i.e. f_x and f_y represent the focal length divided by the pixel width resp. height, (u, v) represents the principal point and s is a factor which is zero in the absence of skew).

A practical way to obtain the calibration parameters from constraints on the intrinsic camera parameters is through application of the concept of the absolute quadric [9]. In space, exactly one degenerate quadric of planes exists which has the property to be invariant under all rigid transformations. In a metric frame it is represented by the following 4×4 symmetric rank 3 matrix $\Omega = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix}$. If $\mathbf{T}$ transforms points $M \to \mathbf{T}M$ (and thus $\mathbf{P} \to \mathbf{P}\mathbf{T}^{-1}$), then it transforms $\Omega \to \mathbf{T}\Omega\mathbf{T}^\top$ (which can be verified to yield Ω when $\mathbf{T}$ is a similarity transformation). The projection of the absolute quadric in the image yields the intrinsic camera parameters independent of the chosen projective basis[2]:

$$\mathbf{K}_i\mathbf{K}_i^\top \propto \mathbf{P}_i\Omega\mathbf{P}_i^\top \tag{3}$$

where $\propto$ means equal up to an arbitrary non-zero scale factor. Therefore constraints on the internal camera parameters in $\mathbf{K}_i$ can be translated to constraints on the absolute quadric. If enough constraints are at hand, only one quadric will satisfy them all, i.e. the *absolute quadric*. At that point the scene can be transformed to the metric frame (which brings Ω to its canonical form).

Equation 3 can be used to obtain the metric calibration from the projective one. A more detailed description of this approach can be found in [9].

2.3 Dense Correspondences

At this point we dispose of a sparse metric reconstruction. Only a few salient points are reconstructed. Obtaining a dense reconstruction could be achieved by interpolation, but in practice this does not yield satisfactory results. Often some salient features are missed during the corner matching and will therefore not appear in the reconstruction. If for example the corner of the roof is missing, this could result in a whole part of the roof missing when using interpolation.

These problems can be avoided by using algorithms which estimate correspondences for almost every point in the images. At this point algorithms can be used which were developed for calibrated 3D systems like stereo rigs. Since we have computed the projective calibration between successive image pairs we can exploit the epipolar constraint that restricts the correspondence search to a 1-d search range. In particular it is possible to remap the image pair to standard geometry where the epipolar lines coincide with the image scan lines [4]. The correspondence search is then reduced to a matching of the image points along each image scanline. 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 exploited. These constraints are used to guide the correspondence towards the most probable scanline match using a dynamic programming scheme [3]. The most recent algorithm [5] improves the accuracy by using a multibaseline approach.

---

[2] Using Equation 2 this can be verified for a metric basis. Transforming $\mathbf{P} \to \mathbf{P}\mathbf{T}^{-1}$ and $\Omega \to \mathbf{T}\Omega\mathbf{T}^\top$ will not change the projection.

Figure 2: Images of the Arenberg castle which were used to generate the 3D model.

Figure 3: Shaded view with cameras (left) and textured view (right) of the 3D model.

2.4 Building the Model

Once a dense correspondence map and the metric camera parameters have been estimated, dense surface depth maps are computed using depth triangulation. The 3D model surface is constructed of triangular surface patches with the vertices storing the surface geometry and the faces holding the projected image color in texture maps. The texture maps add very much to the visual appearance of the models and augment missing surface detail.

The model building process is at present restricted to partial models computed from single view points and work remains to be done to fuse different view points. Since all the views are registered into one metric framework it is possible to fuse the depth estimate into one consistent model surface [4].

Sometimes it is not possible to obtain a single metric framework for large objects like buildings since one may not be able to record images continuously around it. In that case the different frameworks have to be registered to each other. This will be done using available surface registration schemes [2].

3 Experiments

The approach has been used on a wide variety of image sequences. Here some results are given of two images sequences filmed near our department. Figure 2 shows some images of the video sequence which was used to obtain a 3D model of a part of the Arenberg castle in Leuven. The sequence was taken by freely moving a hand held camera. The complete sequence was used to obtain the calibration, while the dense reconstruction is still restricted to what is seen from some reference view (2.5D). The 3D reconstruction is stored as a textured wire-frame VRML model. A few perspective views of this model can be seen in Figure 3. The left view shows the estimated camera viewpoints (little piramids) and the shaded surface view, the right a different textured view. A more quantitative evaluation was obtained by measuring angles in the reconstructed scene between parallel lines (1.8 ± 1.1 degrees) and orthogonal lines (89.7 ± 1.4 degrees). These results confirm the good metric calibration obtained by the method.

4 Conclusion

An automatic 3D scene modelling technique was discussed that is capable of building models from uncalibrated image sequences. The technique is able to extract metric 3D models without any prior knowledge about the scene or the camera. The calibration is obtained by assuming a rigid scene and some constraints on the intrinsic camera parameters (e.g. square pixels).

Work remains to be done to get more complete models by fusing the partial 3D reconstructions. This will also increase the accuracy of the models and eliminate artefacts at the occluding boundaries. For this we can rely on work already done for calibrated systems.

Acknowledgements

We would like to thank Andrew Zisserman and the rest of the VANGUARD team in Oxford for their contributions to the software used for this work. A specialization grant from the Flemish Institute for Scientific Research in Industry (IWT) and the financial support from the EU ACTS project AC074 'VANGUARD' are also gratefully acknowledged.

References

[1] P. Beardsley, P. Torr and A. Zisserman, "3D Model Acquisition from Extended Image Sequences", *Proc. ECCV'96*, Cambridge, UK.

[2] Y. Chen and G. Medioni, "Object Modeling by Registration of Multiple Range Images", *Proc. Int. Conf. on Robotics and Automation*, 1991, Sacramento CA, USA.

[3] L. Falkenhagen, "Hierarchical Block-Based Disparity Estimation Considering Neighbourhood Constraints". International Workshop on SNHC and 3D Imaging, 1997, Rhodes, Greece.

[4] R. Koch, "Automatische Oberflachenmodellierung starrer dreidimensionaler Objekte aus stereoskopischen Rundum-Ansichten", PhD thesis, University of Hannover, 1996.

[5] R. Koch, M. Pollefeys and L. Van Gool, "Multi Viewpoint Stereo from Uncalibrated Video Sequences", *Proc. ECCV'98*, Freiburg, Germany.

[6] M. Pollefeys, R. Koch, M. Vergauwen and L. Van Gool, Virtualizing Archaeological Sites, *Proceedings Virtual Systems and MultiMedia (VSMM'98)*.

[7] M. Pollefeys, R. Koch, M. Vergauwen and L. Van Gool, Metric 3D Surface Reconstruction from Uncalibrated Image Sequences, *Proc. SMILE Workshop (post-ECCV'98)*, LNCS 1506, pp.138-153, Springer-Verlag, 1998.

[8] M. Pollefeys and L. Van Gool, "A Stratified Approach to Metric Self-Calibration", *Proc. CVPR'97*, San Juan, Puerto Rico.

[9] M. Pollefeys, R. Koch and L. Van Gool, "Self-Calibration and Metric 3D reconstruction in Spite of Varying and Unknwon Internal Parameters", *Proc. ICCV'98*, Bombay, India.