AUTOMATIC 3D MODELING FROM IMAGE SEQUENCES

Marc Pollefeys,[*] Maarten Vergauwen and Luc Van Gool
ESAT-PSI, K.U.Leuven, Belgium
Marc.Pollefeys@esat.kuleuven.ac.be

KEY WORDS: 3D reconstruction, structure-from-motion, image sequences, camera (self-)calibration

ABSTRACT

Modeling of three-dimensional (3D) objects from image sequences is a challenging problem and has been a research topic for many years. Important theoretical and algorithmic results were achieved that allow to extract even complex 3D models of scenes from sequences of images. One recent effort has been to reduce the amount of calibration and to avoid restrictions on the camera motion. In this contribution an approach is described which achieves this goal by combining state-of-the-art algorithms for uncalibrated projective reconstruction, self-calibration and dense correspondence matching.

1 INTRODUCTION

Obtaining 3D models from objects is an ongoing research topic. A few years ago the main applications were robot guidance and visual inspection. Nowadays however the emphasis is shifting. There is more and more demand for 3D models in computer graphics, virtual reality and communication. This results in a change in emphasis for the requirements. The visual quality becomes one of the main points of attention. The acquisition conditions and the technical expertise of the users in these new application domains can often not be matched with the requirements of existing systems. These require intricate calibration procedures every time the system is used. There is an important demand for flexibility in acquisition. Calibration procedures should be absent or restricted to a minimum. Additionally, the existing systems are often built around specialized hardware (e.g. laser range scanners or stereo rigs) resulting in a high cost for these systems. Many new applications however require robust low cost acquisition systems. This stimulates the use of consumer photo- or video cameras.

Other researchers have presented systems for extracting 3D shape and texture from image sequences acquired with a freely moving camera. The approach of Tomasi and Kanade(1992) used an affine factorization method to extract 3D from image sequences. An important restriction of this system is the assumption of orthographic projection. Debevec et al.(1996) proposed a system that starts from an approximate 3D model and camera poses and refines the model based on images. View dependent texturing is used to enhance realism. The advantage is that only a restricted number of images are required. On the other hand a preliminary model must be available and the geometry should not be too complex.

In this paper we present a system which retrieves a 3D surface model from a sequence of images taken with off-the-shelf consumer cameras. The user acquires the images by freely moving the camera around the object. Neither the camera motion nor the camera settings have to be known. The obtained 3D model is a scaled version of the original object (i.e. a *metric* reconstruction), and the surface texture is obtained from the image sequence as well. Our system uses full perspective cameras and does not require prior models nor calibration. The complete system combines state-of-the-art algorithms to solve the different subproblems: *projective reconstruction*, *self-calibration* and *dense depth estimation*.

Projective Reconstruction: It has been shown by Faugeras (1992) and Hartley et al. (1992) that a reconstruction up to an arbitrary projective transformation was possible from an uncalibrated image sequence. Since then a lot of effort has been put in reliably obtaining accurate estimates of the projective calibration of an image sequence. Robust algorithms were proposed to estimate the fundamental matrix from image pairs, e.g. Torr (1995) or Zhang et al. (1995). Based on this, algorithms which sequentially retrieves the projective calibration of a complete image sequence have been developed, e.g. Beardsley et al. (1997).

Self-Calibration: Since a projective calibration is not sufficient for many applications, researchers tried to find ways to automatically upgrade projective calibrations to metric (i.e. Euclidean up to scale). Typically, it is assumed that the same camera is used throughout the sequence and that the intrinsic camera parameters are constant. This proved a difficult problem and many researchers have worked on it (Faugeras et al., 1992, Hartley, 1993, Pollefeys and Van Gool, 1999, Triggs, 1997). One of the main problems is that critical motion sequences exist for which self-calibration does not result

[*] Marc Pollefeys is a postdoctoral fellow of the Fund for Scientific Research - Flanders (Belgium) (F.W.O. - Vlaanderen)

in a unique solution (Sturm, 1997). Recently a more pragmatic approach(Pollefeys et al, 1999a) which assumes that some parameters are (approximately) known but which allows others to vary was proposed. Therefore this approach can deal with zooming/focusing cameras.

Dense Depth Estimation: Since the calibration of the image sequence can be estimated, stereoscopic triangulation techniques between image correspondences can be used to estimate depth. The difficult part in stereoscopic depth estimation is to find dense correspondence maps between the images. The correspondence problem is facilitated by exploiting constraints derived from the calibration and from some assumptions about the scene. An approach that combines local image correlation methods with a dynamic programming approach to constrain the correspondence search is used (Cox et al., 1996, Koch, 1996, Falkenhagen, 1997). To obtain a better accuracy a multi-view approach was developed (Koch et al., 1998).

This paper is organized as follows: In section 2 a general overview of the system is given. In the subsequent sections the different steps are explained in more detail: projective reconstruction (section 3), self-calibration (section 4), dense matching (section 5) and model generation (section 6). Section 7 concludes the paper.

2 OVERVIEW OF THE METHOD

The presented system gradually retrieves more information about the scene and the camera setup. The first step is to relate the different images. This is done pairwise by retrieving the epipolar geometry. An initial reconstruction is then made for the first two images of the sequence. For the subsequent images the camera pose is estimated in the projective frame defined by the first two cameras. For every additional image that is processed at this stage, the interest points corresponding to points in previous images are reconstructed, refined or corrected. Therefore it is not necessary that the initial points stay visible throughout the entire sequence. The result of this step is a reconstruction of typically a few hundred to a few thousand interest points and the (projective) pose of the camera. The reconstruction is only determined up to a projective transformation.

The next step consist of restricting the ambiguity of the reconstruction to a metric one. In a projective reconstruction not only the scene, but also the camera is distorted. Since the algorithm deals with unknown scenes, it has no way of identifying this distortion in the reconstruction of the scene. Although the camera is also assumed to be unknown, some constraints on the intrinsic camera parameters (e.g. rectangular or square pixels, constant aspect ratio, principal point in the middle of the image, ...) can often still be assumed. A distortion on the camera mostly results in the violation of one or more of these constraints. A metric reconstruction/calibration is obtained by transforming the projective reconstruction until all the constraints on the cameras intrinsic parameters are satisfied.

At this point the system effectively disposes of a calibrated image sequence. The relative position and orientation of the camera is known for all the viewpoints. This calibration facilitates the search for corresponding points and allows us to use a stereo algorithm that was developed for a calibrated system and which allows to find correspondences for most of the pixels in the images. From these correspondences the distance from the points to the camera center can be obtained through triangulation. These results are refined and completed by combining the correspondences from multiple images.

Finally, a dense metric 3D surface model is obtained by approximating the depth map with a triangular wireframe. The texture is obtained from the images and mapped onto the surface.

In figure 1 an overview of the system 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. Throughout the rest of this paper the different steps of the method will be explained in more detail.

3 PROJECTIVE RECONSTRUCTION

At first the images are completely unrelated. The only assumption is that the images form a sequence in which consecutive images do not differ too much. Therefore the local neighborhood of image points originating from the same scene point should look similar if images are close in the sequence. This allows for automatic matching algorithms to retrieve correspondences. The approach taken to obtain a projective reconstruction is very similar to the one proposed by Beardsley et al (1997).

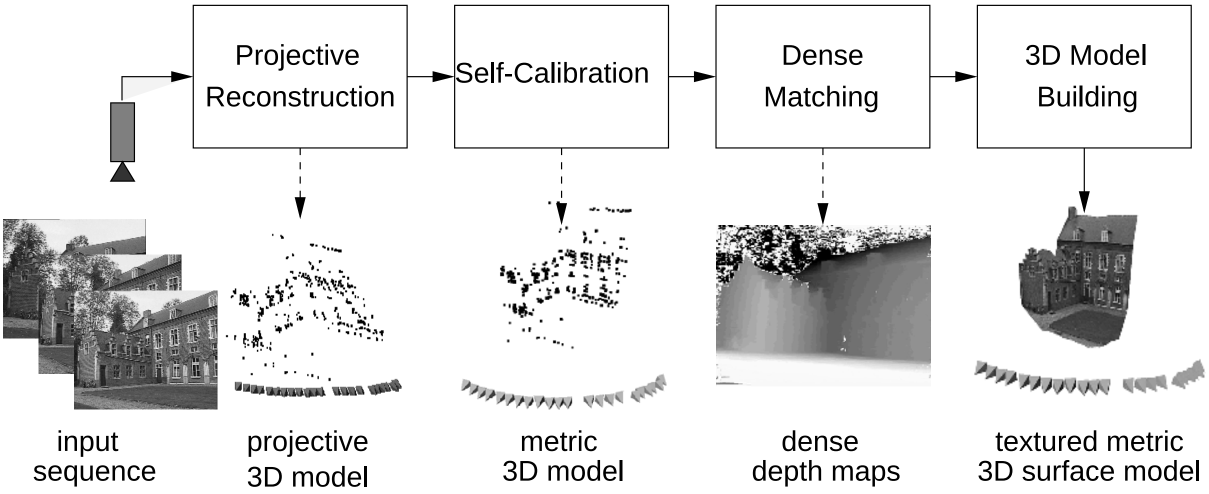

Figure 1: Overview of the system (the cameras are represented by little pyramids, the results of the dense matching are accumulated in dense depth maps where light means close and dark means far).

3.1 Relating the Images

It is not feasible to compare every pixel of one image with every pixel of the next image. It is therefore necessary to reduce the combinatorial complexity. In addition not all points are equally well suited for automatic matching. The local neighborhoods of some points contain a lot of intensity variation and are therefore easy to differentiate from others. The Harris corner detector (Harris and Stephens, 1988) is used to select a set of such points. Correspondences between these image points need to be established through a matching procedure.

Matches are determined through normalized cross-correlation of the intensity values of the local neighborhood. Since images are supposed not to differ too much, corresponding points can be expected to be found back in the same region of the image. Therefore at first only interest points which have similar positions are considered for matching. When two points are mutual best matches they are considered as potential correspondences.

Since the epipolar geometry describes the complete geometry relating two views, this is what should be retrieved. Computing it from the set of potential matches through least squares does in general not give satisfying results due to its sensitivity to outliers. Therefore a robust approach should be used. Our system incorporates the RANSAC (Fischler and Bolles, 1981) approach implemented by Torr (1995). It consist of randomly selecting a minimal set of matches (i.e. 7 for the fundamental matrix) and verifying the consistence of the other matches with the obtained solution. This procedure is repeated until a solution is obtained with sufficient support. Once the epipolar geometry has been retrieved, one can start looking for more matches to refine this geometry. In this case the search region is restricted to a few pixels around the epipolar lines.

3.2 Initial Reconstruction

The two first images of the sequence are used to determine a reference frame. The world frame is aligned with the first camera. The second camera is chosen so that the epipolar geometry corresponds to the retrieved $\mathbf{F}_{12}$, see (Pollefeys, 1999) for more details.

$$\begin{array}{rcllcll} \mathbf{P}_1 & = & [& \mathbf{I}_{3\times3} & | & 0 &] \\ \mathbf{P}_2 & = & [& [\mathsf{e}_{12}]_\times \mathbf{F}_{12} + \mathsf{e}_{12}\mathtt{l}^\top & | & \sigma\mathsf{e}_{12} &] \end{array} \quad (1)$$

where $[\mathsf{e}_{12}]_\times$ indicates the vector product with e_{12}. Equation 1 is not completely determined by the epipolar geometry (i.e. $\mathbf{F}_{12}$ and e_{12}), but has 4 more degrees of freedom (i.e. l_X, l_Y, l_Z, σ). $\mathtt{l} = [l_X l_Y l_Z]^\top$ determines the position of the reference plane (i.e. the plane at infinity in an affine or metric frame) and σ determines the global scale of the reconstruction. To avoid some problems during the reconstruction (due to the violation of the quasi-Euclidean assumption), it is recommended to determine l_X, l_Y, l_Z in such a way that the plane at infinity does not cross the scene. This is achieved by selecting l_X, l_Y, l_Z so that all points are reconstructed in front of the camera. Since there is no way to determine the global scale from the images, σ can arbitrarily be chosen to $\sigma = 1$.

Once the cameras have been fully determined the matches can be reconstructed through triangulation. The optimal method for this is given in (Hartley and Sturm, 1997). This gives us a preliminary reconstruction.

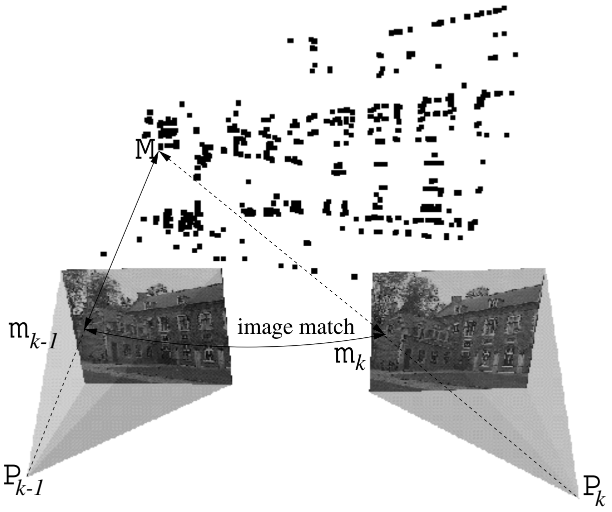

Figure 2: Image matches (m_{k-1}, m_k) are found as described before. Since the image points, m_{k-1}, relate to object points, M_k, the pose for view k can be computed from the inferred matches (M, m_k).

3.3 Adding a View

For every additional view the pose towards the pre-existing reconstruction is determined, then the reconstruction is updated. This is illustrated in figure 2. The first steps consists of finding the epipolar geometry as described in Section 3.1. Then, from the image matches which correspond to already reconstructed points, 2D-3D matches are inferred. These are used to compute the projection matrix $\mathbf{P}_k$. This is done using a robust procedure similar to the one for retrieving the epipolar geometry. In this case a minimal sample of 6 matches is needed to compute $\mathbf{P}_k$. Once $\mathbf{P}_k$ has been determined the projection of already reconstructed points can be predicted. This allows to find some additional matches to refine the estimation of $\mathbf{P}_k$. This means that the search space is gradually reduced from the full image to the epipolar line to the predicted projection of the point. Once the camera projection matrix has been determined the reconstruction is updated. This consists of refining, correcting or deleting already reconstructed points and initializing new points for new matches. This procedure only relates the image to the previous image. In fact it is implicitly assumed that once a point gets out of sight, it will not come back. Although this is true for many sequences, it is certainly not always the case. Therefore, in some cases it can be interesting to adapt the scheme so that more views are matched with the new view (Pollefeys, 1999, Koch et al., 1999).

Once this procedure has been repeated for all the images, one disposes of camera poses for all the views and the reconstruction of the interest points. In the further modules mainly the camera calibration is used. The reconstruction itself is used to obtain an estimate of the disparity range for the dense stereo matching.

4 UPGRADING THE RECONSTRUCTION TO METRIC

The reconstruction obtained as described in the previous paragraph is only determined up to an arbitrary projective transformation. This might be sufficient for some robotics or inspection applications, but certainly not for visualization or metrology. The system uses a flexible self-calibration approach (Pollefeys et al, 1999a, Pollefeys, 1999) to restrict the ambiguity on the reconstruction to metric (i.e. Euclidean up to scale). This approach allows the intrinsic camera parameters to vary during the acquisition. This feature is especially useful when the camera is equipped with a zoom or with auto-focus.

It is outside the scope of this paper to discuss this method in detail. The general concept consist of translating constraints on the intrinsic camera parameters to constraints on the absolute conic. Once this special conic is identified, it can be used as a calibration pattern to upgrade the reconstruction to metric. Some reconstructions *before* and *after* the self-calibration stage are shown. The left part of figure 3 gives the reconstruction before self-calibration. Therefore it is only determined up to an arbitrary projective transformation and metric properties of the scene can not be observed from this representation. The right part shows the result after self-calibration. At this point the reconstruction has been upgraded to metric.

5 DENSE DEPTH ESTIMATION

In the previous steps only a few scene points were reconstructed. Obtaining a dense reconstruction could be achieved by interpolation, but in practice this does not yield satisfactory results. Small surface details would never be reconstructed

Figure 3: Reconstruction before (left) and after (right) self-calibration.

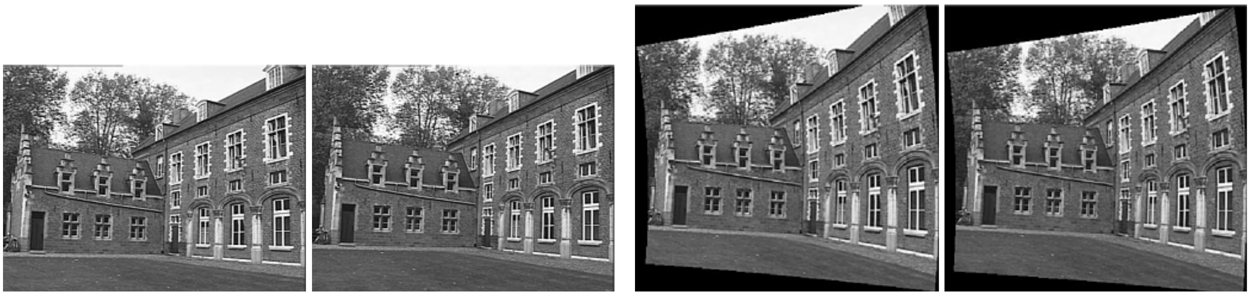

Figure 4: Original image pair (left) and rectified image pair (right).

in this way. Additionally, some important features are often missed during the corner matching and would therefore not appear in the reconstruction. These problems can be avoided by using algorithms which estimate correspondences for almost every point in the images. Because the reconstruction was upgraded to metric, algorithms that were developed for calibrated stereo rigs can be used.

5.1 Rectification

Since the calibration between successive image pairs was computed, the epipolar constraint that restricts the correspondence search to a 1-D search range can be exploited. Image pairs are warped so that epipolar lines coinciding with the image scan lines. The correspondence search is then reduced to a matching of the image points along each image scan-line. This results in a dramatic increase of the computational efficiency of the algorithms by enabling several optimizations in the computations.

For some motions (i.e. when the epipole is located in the image) standard rectification based on planar homographies is not possible and a more advanced procedure should be used. The approach used in the presented system was proposed in (Pollefeys et al. 1999b). The method combines simplicity with minimal image size and works for all possible motions. The key idea is to use polar coordinates with the epipole as origin. A minima image size is achieved by computing the angle between two consecutive epipolar lines to have the worst case pixel on the line preserve its area. figure 4 shows an image pair and the associated rectified image pair.

5.2 Dense Stereo Matching

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 scan-line match using a dynamic programming scheme (Falkenhagen, 1997).

For dense correspondence matching a disparity estimator based on the dynamic programming scheme (Cox et al., 1996), is employed that incorporates the above mentioned constraints. It operates on rectified image pairs where the epipolar lines coincide with image scan lines. The matcher searches at each pixel in the first image for maximum normalized cross correlation in the other image by shifting a small measurement window (kernel size 5x5 to 7x7 pixel) along the corresponding scan line. The selected search step size (usually 1 pixel) determines the search resolution. Matching ambiguities are resolved by exploiting the ordering constraint in the dynamic programming approach (Koch, 1996). The algorithm was further adapted to employ extended neighborhood relationships and a pyramidal estimation scheme to reliably deal with very large disparity ranges of over 50% of image size (Falkenhagen, 1997). This algorithm that was at first developed for calibrate stereo rigs could easily be used for our purposes since at this stage the necessary calibration information had already been retrieved from the images.

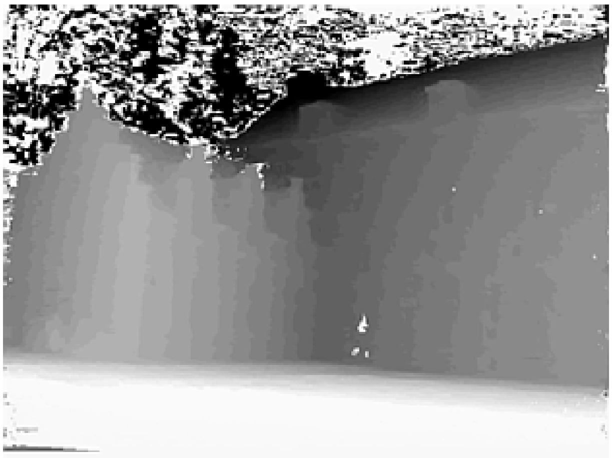

Figure 5: Dense depth map (light means near and dark means far).

5.3 Multi View Matching

The pairwise disparity estimation allows to compute image to image correspondence between adjacent rectified image pairs, and independent depth estimates for each camera viewpoint. An optimal joint estimate is achieved by fusing all independent estimates into a common 3D model. The fusion can be performed in an economical way through controlled correspondence linking. The approach utilizes a flexible multi viewpoint scheme which combines the advantages of small baseline and wide baseline stereo (Koch et al., 1998). The result of this procedure is a very dense depth map. Most occlusion problems are avoided by linking correspondences from up and down the sequence. An example of such a very dense depth map is given in figure 5.

6 BUILDING THE MODEL

The dense depth maps as computed by the correspondence linking must be approximated by a 3D surface representation suitable for visualization. For this purpose a triangular mesh is overlaid on top of the depth map and the triangles are backprojected in space according to the depth of the vertices. The original image itself can be used as a texture to enhance realism. Note that in this case the problem of registering the texture with the 3D model is trivial. To avoid highlights and other artefacts which could be present in the reference image a robust texture can be build-up through the same scheme as was used to refine depth in the previous section. This is described more in detail in (Koch et al., 1998).

An example of the resulting model can be seen in figure 6. Some more views of the reconstruction are given in figure 7. To further illustrate the flexibility of the system a second example is given. The 5 images seen in figure 8 were taken with a simple photocamera and transfered to a PhotoCD. Feature points were extracted and matched automatically between these images and the calibration was obtained as described in this paper. Next, a full surface model was computed from this. These results are illustrated in figure 9. Due to the flexibility of the system, it could for example also be used to reconstruct scenes from pre-existing video (Pollefeys, 1999) or to obtain the calibration required to construct plenoptic models (Koch et al., 1999).

7 CONCLUSION

An automatic 3D scene modeling 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). Future research will try to deal with more widely separated views and to obtain a better accuracy through maximum likelihood estimation. Work also remains to be done to get more complete models by fusing the partial 3D reconstructions.

ACKNOWLEDGMENTS

We wish to acknowledge the financial support of the Belgian IUAP 24/02 ’ImechS’ project and of the ITEA99002 BEYOND project that is supported by the flemish district - IWT.

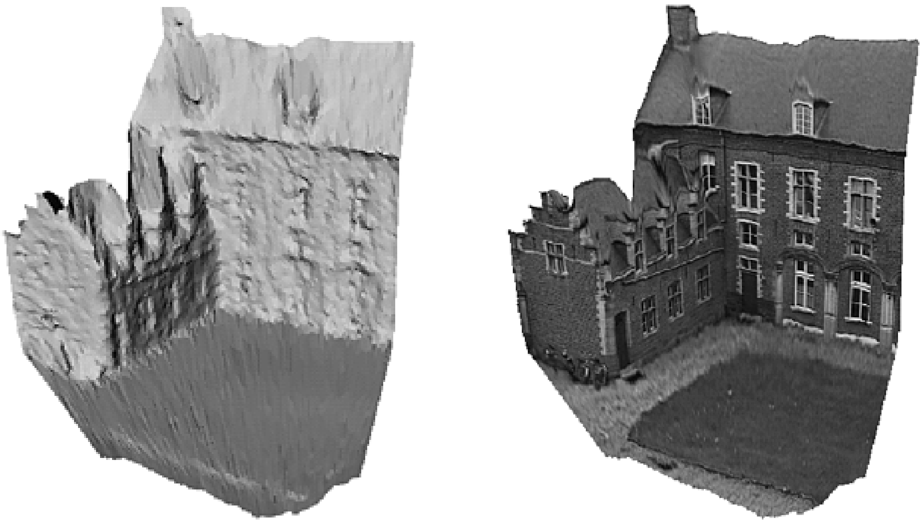

Figure 6: 3D surface model obtained automatically from an uncalibrated image sequence, shaded (left), textured (right).

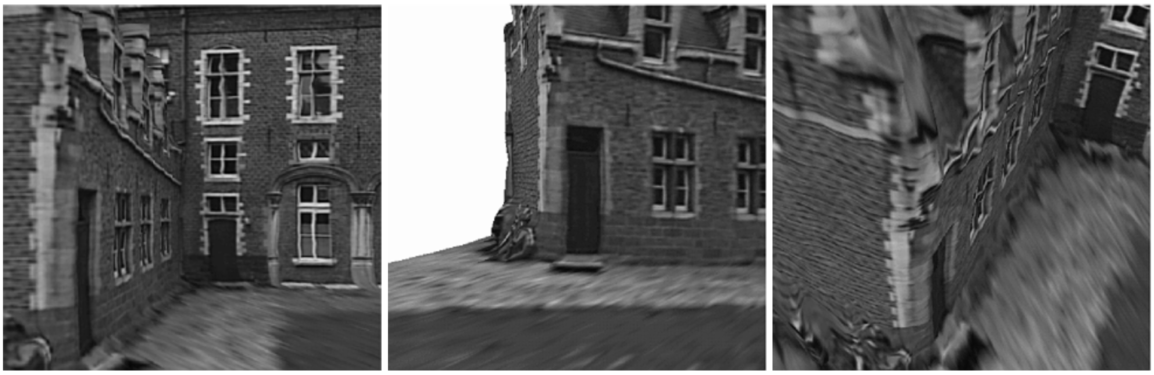

Figure 7: Some detailed views of the reconstructed castle model.

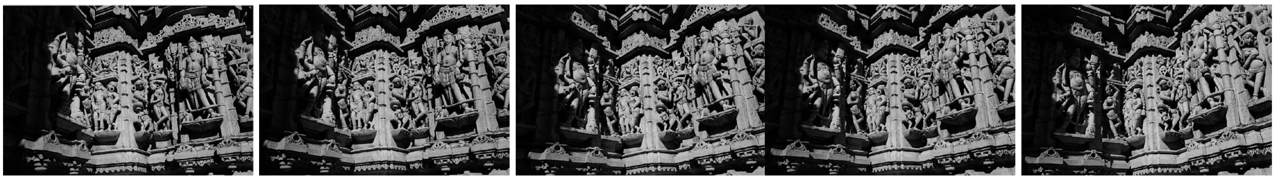

Figure 8: *Photographs which were used to generate a 3D model of a part of a Jain temple of Ranakpur.*

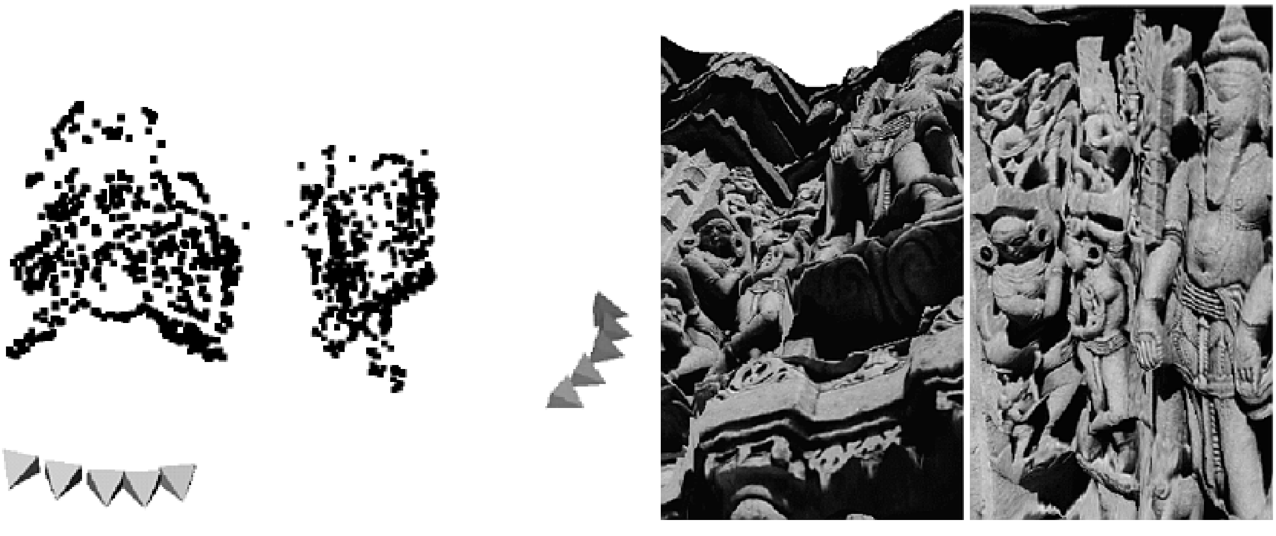

Figure 9: *Reconstruction of interest points and cameras (left), two detail views of the reconstructed model (right).*

REFERENCES

Beardsley, P., Zisserman, A., and Murray, D., 1997. Sequential Updating of Projective and Affine Structure from Motion, Int. J. of Computer Vision, 23(3), pp. 235-259.

Cox, I., Hingorani, S., and Rao, S., 1996. A Maximum Likelihood Stereo Algorithm, Computer Vision and Image Understanding, 63(3).

Debevec, P., Taylor, C., and Malik, J., 1996. Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach, Proc. ACM Siggraph.

Falkenhagen, L., 1997. Hierarchical Block-Based Disparity Estimation Considering Neighbourhood Constraints. Proc. Int. Workshop on SNHC and 3D Imaging.

Faugeras, O., 1992. What can be seen in three dimensions with an uncalibrated stereo rig, Computer Vision - ECCV'92, LNCS, Vol. 588, Springer-Verlag, pp. 563-578.

Faugeras, O., Luong, Q.-T., and Maybank, S., 1992. Camera self-calibration: Theory and experiments, Computer Vision - ECCV'92, LNCS, Vol. 588, Springer-Verlag, pp. 321-334.

Fischler, M., and Bolles, R., 1981. RANdom SAmpling Consensus: a paradigm for model fitting with application to image analysis and automated cartography, Commun. Assoc. Comp. Mach., 24, pp.381-95.

Harris, C., and Stephens, M., 1988. A combined corner and edge detector, Fourth Alvey Vision Conference, pp.147-151.

Hartley, R., Gupta, R., and Chang, T., 1992. Stereo from Uncalibrated Cameras, Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.761-764.

Hartley, R., 1994. Euclidean reconstruction from uncalibrated views, in : J.L. Mundy, A. Zisserman, and D. Forsyth (eds.), Applications of Invariance in Computer Vision, LNCS, Vol. 825, Springer-Verlag, pp. 237-256.

Hartley, R., and Sturm, P., 1997. Triangulation, Computer Vision and Image Understanding, 68(2), pp.146-157.

Koch, R., 1996, Automatische Oberflachenmodellierung starrer dreidimensionaler Objekte aus stereoskopischen Rundum-Ansichten, PhD thesis, University of Hannover, Germany, 1996 also published as Fortschritte-Berichte VDI, Reihe 10, Nr.499, VDI Verlag, 1997.

Koch, R., Pollefeys, M., and Van Gool, L., 1998. Multi Viewpoint Stereo from Uncalibrated Video Sequences. Computer Vision - ECCV'98, LNCS, 1406, Springer-Verlag, pp.55-71.

Koch, R., Pollefeys, M., Heigl, B., Van Gool, L., and Niemann, H., 1999. Calibration of Hand-held Camera Sequences for Plenoptic Modeling, Proc. Int. Conf. on Computer Vision, IEEE Computer Society Press, pp.585-591.

Pollefeys, M., and Van Gool, L., 1999. Stratified self-calibration with the modulus constraint, IEEE Transactions on Pattern Analysis and Machine Intelligence, 21 (8), pp.707-724.

Pollefeys, M., 1999. Self-calibration and metric 3D reconstruction from uncalibrated image sequences, Ph.D. Thesis, ESAT-PSI, K.U.Leuven.

Pollefeys, M., Koch, R., and Van Gool, L., 1999 (a). Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters, Int. J. of Computer Vision, 32(1), pp.7-25.

Pollefeys, M., Koch, R., and Van Gool, L., 1999 (b). A simple and efficient rectification method for general motion, Proc. Int. Conf. on Computer Vision, IEEE Computer Society Press, pp.496-501.

Sturm, P., 1997. Critical Motion Sequences for Monocular Self-Calibration and Uncalibrated Euclidean Reconstruction, Proc. Conf. on Computer Vision and Pattern Recognition, IEEE Computer Soc. Press, pp. 1100-1105.

Tomasi, C. and Kanade, T., 1992. Shape and motion from image streams under orthography: A factorization approach, Int. J. of Computer Vision, 9(2), pp.137-154.

Torr, P., 1995. Motion Segmentation and Outlier Detection, PhD Thesis, Dept. of Engineering, Univ. of Oxford.

Triggs, B., 1997. The Absolute Quadric, Proc. Conf. on Computer Vision and Pattern Recognition, IEEE Computer Soc. Press, pp. 609-614.

Zhang, Z., Deriche, R., Faugeras, O., and Luong, Q.-T., 1995. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry, Artificial Intelligence J., pp.87-119.