3D CAPTURE OF ARCHAEOLOGY AND ARCHITECTURE WITH A HAND-HELD CAMERA

Marc Pollefeys[a,*]**, Luc Van Gool**[b]**, Maarten Vergauwen**[b]**, Kurt Cornelis**[b]**, Frank Verbiest**[b]**, Jan Tops**[b]
[a] Dept. of Computer Science, University of North Carolina – Chapel Hill,
[b] Center for Processing of Speech and Images, K.U.Leuven
marc@cs.unc.edu

Working Group III/V

KEY WORDS: 3D modeling, video sequences, image sequences, archaeology, architectural conservation.

ABSTRACT

In this paper we present an automated processing pipeline that, from a sequence of images, reconstructs a 3D model. The approach is particularly flexible as it can deal with a hand-held camera without the need for an a priori calibration or explicit knowledge about the recorded scene. In a fist stage features are extracted and tracked throughout the sequence. Using robust statistics and multiple view relations the 3D structure of the observed features and the camera motion and calibration are computed. In a second stage stereo matching is used to obtain a detailed estimate of the geometry of the observed scene. The presented approach integrates state-of-the-art algorithms developed in computer vision, computer graphics and photogrammetry. Due to its flexibility during image acqusition, this approach is particularly well suited for application in the field of archaeology and architectural conservation.

1 INTRODUCTION

In the context of archaeology, measurement and documentation are very important, not only to record endangered archaeological monuments or sites, but also to record the excavations themselves. Archaeology is one of the sciences were annotations and precise documentation are most important because evidence is destroyed during the normal progress of work. Therefore, on most archaeological sites a large amount of time is spent taking measurements, drawing plans, making notes and taking photographs, etc. Archaeologists are also becoming aware of the possibilities of 3D visualization for dissemination to a wider public. These models can also be used for planning restorations or as digital archives, although many issues still have to be solved here. However, up to now the cost in time and money to generate this type of virtual reconstructions remains prohibitive for most archaeological projects.

The image-based 3D recording approach that we have developed over the last few years (Pollefeys et al., 1999b, Koch et al. 1998, Pollefeys et al., 1999a, Van Meerbergen et al. 2002, Pollefeys et al. 2001a) offers a lot of possibilities. To acquire a 3D reconstruction of an object or a scene it is sufficient to take a number of pictures from different viewpoints. These can be obtained using a photo or video camera. Consecutive pictures should not differ too much so that the computer can automatically identify matching features. It is not necessary for the camera to be calibrated and if necessary the focal length of the camera (zoom, focus) can be changed during acquisition. In principle no additional measurements have to be take in the scene to obtain a 3D model. However, a reference length can be useful to obtain the global scale of the reconstruction. If the absolute localization in world coordinates is required the measurement of 3 reference points might be necessary. The resulting 3D model can be used both for measurements and visualization purposes.

*corresponding author

The remainder of this paper is organized as follows. First our 3D recording approach is presented. Then different archaeological and architectural applications are discussed.

2 3D CAPTURE FROM IMAGES

Starting from a sequence of images the first step consists of recovering the relative motion between consecutive images. This process goes hand in hand with finding corresponding image features between these images (i.e. image points that originate from the same 3D feature). The next step consists of recovering the motion and calibration of the camera and the 3D structure of the features. This process is done in two phases. At first the reconstruction contains a projective skew (i.e. parallel lines are not parallel, angles are not correct, distances are too long or too short, etc.). This is due to the absence of an a-priori calibration. Using a self-calibration algorithm (Pollefeys et al., 1999b) this distortion can be removed, yielding a reconstruction equivalent to the original up to a global scale factor. This *uncalibrated* approach to 3D reconstruction allows much more flexibility in the acquisition process since the focal length and other intrinsic camera parameters do not have to be measured –calibrated– beforehand and are allowed to change during the acquisition.

The reconstruction obtained as described in the previous paragraph only contains a sparse set of 3D points. Although interpolation might be a solution, this yields models with poor visual quality. Therefore, the next step consists in an attempt to match all image pixels of an image with pixels in neighboring images, so that these points too can be reconstructed. This task is greatly facilitated by the knowledge of all the camera parameters which we have obtained in the previous stage. Since a pixel in the image corresponds to a ray in space and the projection of this ray in other images can be predicted from the recovered pose and calibration, the search of a corresponding pixel in other images can be restricted to a single line. Additional

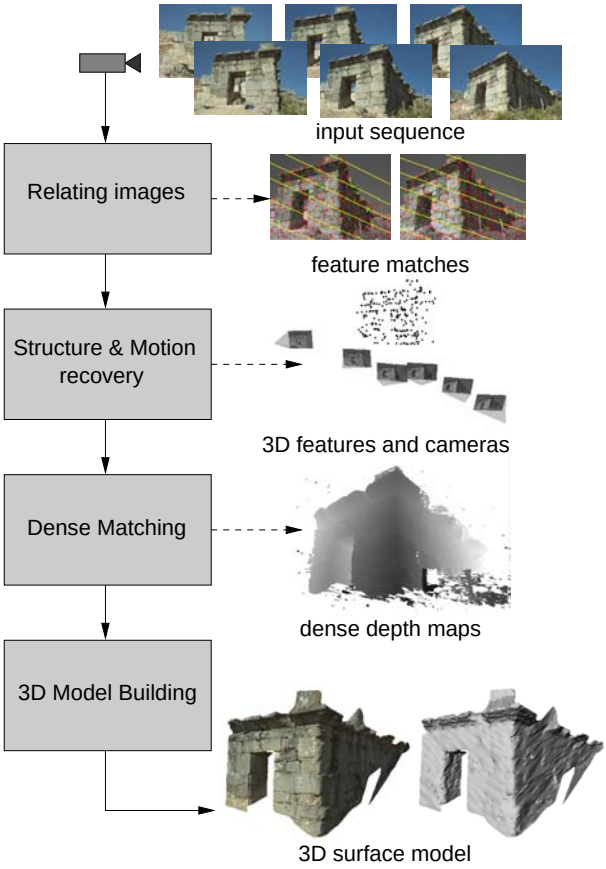


Figure 1: Overview of our image-based 3D recording approach.

constraints such as the assumption of a piecewise continuous 3D surface are also employed to further constrain the search. It is possible to warp the images so that the search range coincides with the horizontal scan-lines. An algorithm that can achieve this for arbitrary camera motion is described in (Pollefeys et al., 1999a). This allows us to use an efficient stereo algorithm that computes an optimal match for the whole scan-line at once (Van Meerbergen et al. 2002). Thus, we can obtain a depth estimate (i.e. the distance from the camera to the object surface) for almost every pixel of an image. By fusing the results of all the images together a complete dense 3D surface model is obtained. The images used for the reconstruction can also be used for texture mapping so that a final photo-realistic result is achieved. The different steps of the process are illustrated in Figure 1. In the following paragraphs the different steps are described in some more detail.

2.1 Relating images

Starting from a collection of images or a video sequence the first step consists of relating the different images to each other. This is not an easy problem. A restricted number of corresponding points is sufficient to determine the geometric relationship or *multi-view constraints* between the images. Since not all points are equally suited for matching or tracking (e.g. a pixel in a homogeneous region), feature points need to be selected (Harris and Stephes, 1988). Depending on the type of image data (i.e. video or still pictures) the feature points are tracked or matched

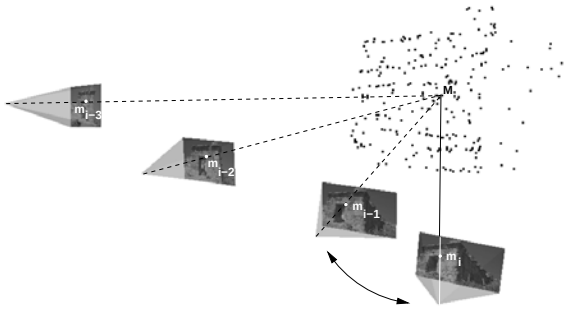


Figure 2: The pose estimation of a new view uses inferred structure-to-image matches.

and a number of potential correspondences are obtained. From these the multi-view constraints can be computed. However, since the correspondence problem is an ill-posed problem, the set of corresponding points can (and almost certainly will) be contaminated with an important number of wrong matches or *outliers*. A traditional least-squares approach will fail and therefore a robust method is used (Torr, 1995, Fischler, 1981). Once the multi-view constraints have been obtained they can be used to guide the search for additional correspondences. These can then be employed to further refine the results for the multi-view constraints.

2.2 Structure and motion recovery

The relation between the views and the correspondences between the features, retrieved as explained in the previous section, will be used to retrieve the structure of the scene and the motion of the camera. Our approach is fully projective so that it does not depend on the initialization. This is achieved by strictly carrying out all measurements in the images, i.e. using reprojection errors instead of 3D errors.

At first two images are selected and an initial projective reconstruction frame is set-up (Faugeras, 1992, Hartley et al., 1992). Matching feature points are reconstructed through triangulation. Features points that are also observed in a third view can then be used to determine the pose of this view in the reference frame defined by the two first views. The initial reconstruction is then refined and extended. By sequentially applying the same procedure the structure and motion of the whole sequence can be computed. The pose estimation procedure is illustrated in Figure 2. These results can be refined through a global least-squares minimization of all reprojection errors. Efficient bundle adjustment techniques (Triggs et al., 2000, Slama, 1980) have been developed for this. Then the ambiguity is restricted to metric through self-calibration (Pollefeys et al., 1999b). Finally, a second bundle adjustment is carried out that takes the camera calibration into account to obtain an optimal estimation of the metric structure and motion.

2.3 Dense surface estimation

To obtain a more detailed model of the observed surface a dense matching technique is used. The structure and motion obtained in the previous steps can be used to constrain

Figure 3: Example of a rectified stereo pair.

the correspondence search. 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 coincide with the image scan lines. For this purpose the rectification scheme proposed in (Pollefeys et al., 1999a) is used. This approach can deal with arbitrary relative camera motion which is not the case for standard homography-based approaches which fail when the epipole is contained in the image. The approach proposed in (Pollefeys et al., 1999a) also guarantees minimal image size. 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. An example of a rectified stereo pair is given in Figure 3. Note that all corresponding points are located on the same horizontal scan-line in both images.

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 (Van Meerbergen et al. 2002). The matcher searches at each pixel in one image for maximum normalized cross correlation in the other image by shifting a small measurement window along the corresponding scan line. The algorithm employs a pyramidal estimation scheme to reliably deal with very large disparity ranges of over 50% of image size. The disparity search range is limited based on the disparities that were observed for the features in the previous reconstruction stage.

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 using a Kalman filter. The fusion can be performed in an economical way through controlled correspondence linking and was discussed more in detail in (Koch et al. 1998). This approach combines the advantages of small baseline and wide baseline stereo. It can provide a very dense depth map by avoiding most occlusions. The depth resolution is increased through the combination of multiple viewpoints and large global baseline while the matching is simplified through the small local baselines.

2.4 Building virtual models

In the previous sections a dense structure and motion recovery approach was explained. This yields all the necessary information to build textured 3D models. The 3D surface is approximated by a triangular mesh to reduce geometric complexity and to tailor the model to the requirements of computer graphics visualization systems. A simple approach consists of overlaying a 2D triangular mesh on top of one of the images and then build a corresponding 3D mesh by placing the vertices of the triangles in 3D space according to the values found in the corresponding depth map. The image itself is used as texture map. If no depth value is available or the confidence is too low the corresponding triangles are not reconstructed. The same happens when triangles are placed over discontinuities. This approach works well on dense depth maps obtained from multiple stereo pairs.

The texture itself can also be enhanced through the multi-view linking scheme. A median or robust mean of the corresponding texture values can be computed to discard imaging artifacts like sensor noise, specular reflections and highlights (Koch et al. 1998).

To reconstruct more complex shapes it is necessary to combine multiple depth maps. Since all depth-maps are located in a single metric frame, registration is not an issue. To integrate the multiple depth maps into a single surface representation, the volumetric technique proposed by Curless and Levoy (Curless and Levoy, 1996) is used.

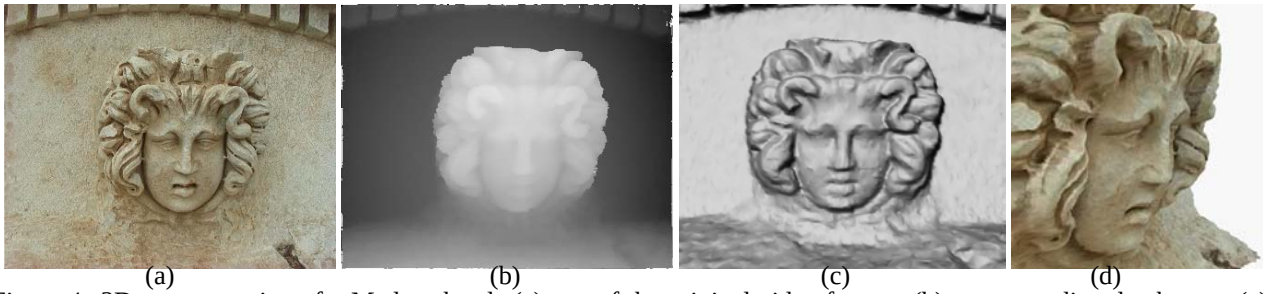
3 APPLICATIONS TO ARCHAEOLOGY

The technique described in the previous section has many applications in the field of archaeology. In this section we will discuss several of these and illustrate them with results that were obtained at Sagalassos. First we will show some results obtained on different types of 3D objects and scenes found at Sagalassos. Then more specific applications of our technique will be discussed.

3.1 Acquiring 3D scenes

The 3D surface acquisition technique that was presented in the previous section, can readily be applied to archaeological field work. The on-site acquisition procedure consists of recording an image sequence of the scene that one desires to reconstruct. To allow the algorithms to yield good results viewpoint changes between consecutive images should not exceed 5 to 10 degrees.

An example is shown in Figure 4. It is a Medusa head which is located on the entablature of a monumental fountain in Sagalassos. The head itself is about $30cm$ across. The 3D model was obtained from a short video sequence. In this case also a single depth map was used to reconstruct the 3D model. Notice that a realistic view is obtained from a viewpoint that is very different from the original viewpoint. The accuracy of the reconstruction should be considered at two levels. Errors on the camera motion and



(a) (b) (c) (d)

Figure 4: 3D reconstruction of a Medusa head. (a) one of the original video frames, (b) corresponding depth map, (c) shaded view of the 3D model and (d) textured view of the 3D model.

calibration computations can result in a global bias on the reconstruction. From the results of the bundle adjustment we can estimate this error to be of the order of $3mm$ for points on the reconstruction. The depth computations indicate that 90% of the reconstructed points have a relative error of less than $1mm$. Note that the stereo correlation uses a 7×7 window which corresponds to a size of $5mm\times5mm$ on the object and therefore the measured depth will typically correspond to the dominant visual feature within that patch.

An important advantage of our approach compared to more interactive techniques (1, 2) is that much more complex objects can be dealt with. Compared to non-image based techniques we have the important advantage that surface texture is directly extracted from the images. This does not only result in a much higher degree of realism, but is also important for the authenticity of the reconstruction. Therefore the reconstructions obtained with this system can also be used as a scale model on which measurements can be carried out or as a tool for planning restorations.

3.2 Recording 3D Stratigraphy

Archaeology is one of the sciences were annotations and precise documentation are most important because evidence is destroyed during work. An important aspect of this is the stratigraphy. This reflects the different layers of soil that correspond to different time periods in an excavated sector. Due to practical limitations this stratigraphy is often only recorded for some slices, not for the whole sector.

Our technique allows a more optimal approach. For every layer a complete 3D model of the excavated sector can be generated. Since the technique only involves taking a series of pictures this does not slow down the progress of the archaeological work. The excavations of an ancient Roman villa at Sagalassos were recorded with our technique. In Fig. 5 a view of the 3D model of the excavation is provided for several layers. The on-site acquisition time was around 1 minute per layer. The reconstructed area is approximately $5m \times 5m$. From the results of the bundle adjustment we can estimate the global error to be of the order of $1cm$ for points on the reconstruction. Similarly the depth computations indicate that the depth error of most of the reconstructed points should be within $1cm$. Note that in this case the correlation window corresponds to approximately $5cm \times 5cm$ in the scene and some small details might thus not appear in the reconstruction. This accuracy

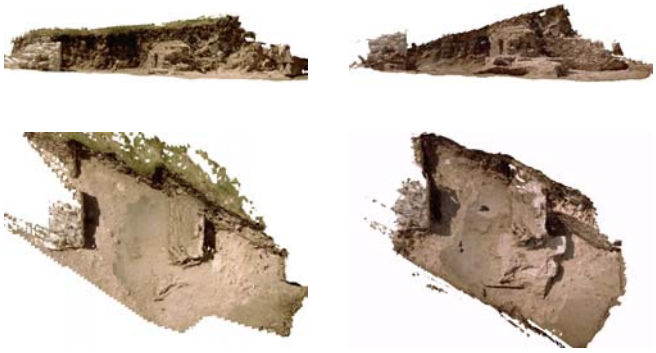


Figure 5: 3D stratigraphy, the excavation of a Roman villa. The front and top view of two different stratigraphics layers is shown.

is more than sufficient to satisfy the requirements of the archaeologists.

To obtain a single 3D representation for each stratigraphic layer, the volumetric integration approach of Curless and Levoy was used (Curless and Levoy, 1996).

3.3 Generating and testing construction hypotheses

The technique proposed in this paper also has a lot to offer for generating and testing construction hypotheses. Due to the ease of acquisition and the obtained level of detail, one could reconstruct every building block separately. The different construction hypotheses can then interactively be verified on a virtual reconstruction site. Registration algorithms (Chen and Medioni, 1991) can even be used to automate this. Fig. 6 shows two segments of a broken column. The whole monument contains 16 columns that were all broken in several pieces by an earthquake. Since each piece can weigh several hundreds of kilograms, physically trying to fit pieces together is a very heavy task. Traditional drawings also do not offer a proper solution.

3.4 Fusion of real and virtual scenes

An interesting possibility is the combination of recorded 3D models with other type of models. In the case of Sagalassos some reconstruction drawings were translated to CAD models (Martens et al., 2000). These were integrated with our models. The result can be seen in Fig. 7. Other models,

Figure 6: Two images of a broken pillar (left) and the orthographic views of the matching surfaces generated from the obtained 3D models (right).

Figure 7: Virtualized landscape of Sagalassos combined with CAD-models of reconstructed monuments

obtained with different 3D acquisition techniques, could also easily be integrated. This reconstruction is available on the Internet (http://www.esat.kuleuven.ac.be/sagalassos/).

Another challenging application consists of seamlessly integrating virtual objects in real video. In this case the ultimate goal is to make it impossible to differentiate between real and virtual objects. Several problems need to be overcome before achieving this goal. Amongst them the rigid registration of virtual objects with the real environment is the most important. This can be done using the computed camera motion (Cornelis et al., 2001). An important difference with the applications discussed in the previous sections is that in this case all frames of the input video sequence have to be processed while for 3D modeling often a sparse set of views is sufficient. Therefore, in this case it is more appropriate to track features from frame to frame.To allow a successful insertion of large virtual objects in an image sequence, the recovered 3D structure should not be distorted. For this purpose it is important to use a camera model that takes non-perspective effects –such as radial distortion– into account and to perform a global least-squares minimization of the reprojection error through a bundle adjustment.

The following example was recorded at Sagalassos in Turkey, where footage of the ruins of an ancient fountain was taken. A large part of the original monument is missing. Based on results of archaeological excavations and architectural studies, it was possible to generate a virtual copy of the missing part. Using the proposed approach the virtual reconstruction could be placed back on the remains of the original monument, at least in the recorded video sequence. In Figure 8 a frame of the augmented video sequence is shown. This sequence was recorded for a documentary by Axell Communication.

Figure 8: Architect contemplating a virtual reconstruction of the nymphaeum of the upper agora at Sagalassos.

4 APPLICATION TO ARCHITECTURAL CONSERVATION

Our approach can also be applied to the recording of architectural monuments. This is especially relevant in the context of conservation. In Fig. 9 a few examples of recorded 3D models are shown. In this case, a number of reference points were measured to obtain an absolute localization and scale. More details on this can be found in (Schouteden et al., 2001).

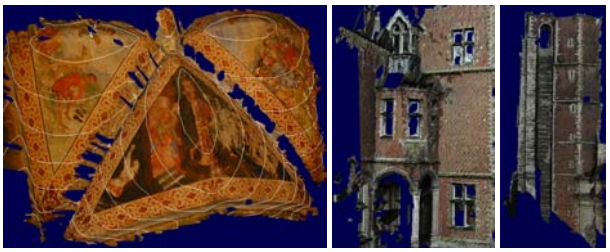


Figure 9: Some examples of parts of architectural monuments that have been recorded using our approach.

5 CONCLUSION

In this paper the application of an image-based 3D recording technique to archaeological field work was presented. The approach utilizes different components that gradually retrieve all the information that is necessary to construct virtual models from images. Automatically extracted features are tracked or matched between consecutive views and multi-view relations are robustly computed. Based on this the projective structure and motion is determined and subsequently upgraded to metric through self-calibration. Bundle-adjustment is used to refine the results. Then, image pairs are rectified and matched using a stereo algorithm and dense and accurate depth maps are obtained by combining measurements of multiple pairs. This technique was successfully applied to the acquisition of virtual models on archaeological sites. There are multiple advantages: the on-site acquisition time is restricted, the construction of the models is automatic and the generated models are realistic.

The technique allows some very promising applications such as 3D stratigraphy recording, the (automatic) generation and verification of construction hypotheses, the 3D reconstruction of scenes based on archive photographs or video footage and integrating virtual reconstructions with archaeological remains in video footage.

Acknowledgment

We would like to thank Prof. Dr. Marc Waelkens and his team for making it possible for us to do experiments at the archaeological site of Sagalassos (Turkey). The partial financial support of the NSF IIS 0237533, FWO project G.0223.01 and the IST-1999-20273 project 3DMURALE are gratefully acknowledged. Kurt Cornelis is research assistant of the Fund for Scientific Research - Flanders (Belgium).

REFERENCES

Y. Chen and G. Medioni, "Object Modeling by Registration of Multiple Range Images", *IEEE. International Conference on Robotics and Automation*, pp 2724-2729, 1991.

K. Cornelis, M. Pollefeys, M. Vergauwen and L. Van Gool, "Augmented Reality from Uncalibrated Video Sequences", In M. Pollefeys, L. Van Gool, A. Zisserman, A. Fitzgibbon (Eds.), *3D Structure from Images - SMILE 2000*, Lecture Notes in Computer Science, Vol. 2018, pp.150-167. Springer-Verlag, 2001.

B. Curless and M. Levoy, "A Volumetric Method for Building Complex Models from Range Images" *Proc. SIGGRAPH '96*, pp. 303–312, 1996.

P. Debevec, C. Taylor and J. Malik, "Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach", *Proc. SIGGRAPH'96*, pp. 11–20, 1996.

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

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

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

R. Hartley, R. Gupta, and T. Chang. "Stereo from uncalibrated cameras". *Proc. Conference Computer Vision and Pattern Recognition*, pp. 761-764, 1992.

R. Koch, M. Pollefeys and L. Van Gool, "Multi Viewpoint Stereo from Uncalibrated Video Sequences". *Proc. European Conference on Computer Vision*, pp.55-71. Freiburg, Germany, 1998.

F. Martens, P. Legrand, J. Legrand, L. Loots and M. Waelkens, "Computer-aided design and archaeology at Sagalassos: methodology and possibilities of reconstructions of archaeological sites", In J. Barcelo, M. Forte and D. Sanders, *Virtual Reality in Archaeology*, ArcheoPress, Oxford (British Archaeological Reports, International Series #843), 205-212, 2000.

PhotoModeler, by Eos Systems Inc., (for more information: `http://www.photomodeler.com`).

M. Pollefeys, R. Koch and L. Van Gool, "A simple and efficient rectification method for general motion", *Proc.ICCV'99 (international Conference on Computer Vision)*, pp.496-501, Corfu (Greece), 1999.

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

M. Pollefeys, M. Vergauwen, F, Verbiest, K. Cornelis, J. Tops and L. Van Gool. "Virtual Models from Video and Vice-Versa, keynote presentation", *Proc. International Symposium on Virtual and Augmented Architecture (VAA01)*, B. Fisher, K. Dawson-Howe, C. O'Sullivan (Eds.), Springer-Verlag, pp.11-22+plate 1, 2001.

J. Schouteden, M. Pollefeys, M. Vergauwen, L. Van Gool, "Image-based 3D acquisition tool for architectural conservation", *Proc. CIPA conference, International Archive of Photogrammetry and Remote Sensing*, 2001.

C. Slama, *Manual of Photogrammetry*, American Society of Photogrammetry, Falls Church, VA, USA, 4th edition, 1980.

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

Triggs, B. 1997. The Absolute Quadric, In *Proc. International Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp.609-614.

B. Triggs, P. McLauchlan, R. Hartley, A. Fiztgibbon, "Bundle Adjustment – A Modern Synthesis", In B. Triggs, A. Zisserman, R. Szeliski (Eds.), *Vision Algorithms: Theory and Practice*, LNCS Vol.1883, pp.298-372, Springer-Verlag, 2000.

G. Van Meerbergen, M. Vergauwen, M. Pollefeys, L. Van Gool, "A Hierarchical Symmetric Stereo Algorithm Using Dynamic Programming", *International Journal of Computer Vision*, Vol. 47, No. 1-3, 2002.