VIDEO-TO-3D

Marc Pollefeys[a,b,*], Luc Van Gool[a], Maarten Vergauwen[a], Kurt Cornelis[a], Frank Verbiest[a], Jan Tops[a]

[a] Center for Processing of Speech and Images, K.U.Leuven
[b] Dept. of Computer Science, University of North Carolina – Chapel Hill,
Marc.Pollefeys@cs.unc.edu

Working Group III/V

KEY WORDS: 3D modeling, video sequences, structure from motion, self-calibration, stereo matching, image-based rendering.

ABSTRACT:

In this contribution we intend to present a complete system that takes a video sequence of a static scene as input and generates visual 3D model. The system can deal with images acquired by an uncalibrated hand-held camera, with intrinsic camera parameters possibly varying during the acquisition. In a first 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. The resulting models are suited for both measurement and visualization purposes.

1 INTRODUCTION

In recent years the emphasis for applications of 3D modeling has shifted from measurements to visualization. New communication and visualization technology have created an important demand for photo-realistic 3D content. In most cases virtual models of existing scenes are desired. This has created a lot of interest for image-based approaches. Applications can be found in e-commerce, real estate, games, post-production and special effects, simulation, etc. For most of these applications there is a need for simple and flexible acquisition procedures. Therefore calibration should be absent or restricted to a minimum. Many new applications also require robust low cost acquisition systems. This stimulates the use of consumer photo- or video cameras. The approach presented in this paper allows to captures photo-realistic virtual models from images. The user acquires the images by freely moving a camera around an object or scene. Neither the camera motion nor the camera settings have to be known a priori. There is also no need for preliminary models. The approach can also be used to combine virtual objects with real video, yielding augmented video sequences.

The approach proposed in this papers builds further on earlier work, e.g. (Pollefeys et al., 2000). Several important improvements were made to the system. To deal more efficiently with video, we have developed an approach that can automatically select key-frames suited for structure and motion recovery. The projective structure and motion recovery stage has been made completely independent of the initialization which avoids some instability problems that occurred with the quasi-Euclidean initialization proposed in (Beardsley et al., 1997). Several optimizations have been implemented to obtain more efficient robust algorithms (Matas and Chum, 2001). To guarantee a maximum likelihood reconstruction at the different levels a state-of-the-art bundle adjustment algorithm was implemented that can be used both at the projective and the Euclidean level. A much more robust linear self-calibration algorithm was obtained by incorporating general a priori knowledge on meaningful values for the camera intrinsics. This allows to avoid most problems related to critical motion sequences (Sturm, 1997) (i.e. some motions do not yield a unique solution for the calibration of the intrinsics) that caused the initial linear algorithm proposed in (Pollefeys et al., 1998) to yield poor results under some circumstances. A solution was also developed for another problem. Previously observing a purely planar scene at some point during the acquisition would have caused uncalibrated approaches to fail. A solution that detects this case and deals with it accordingly has been proposed (Pollefeys et al., 2002). Both correction for radial distortion and stereo rectification have been integrated in a single image resampling pass. This allows to minimize the image degradation. Our processing pipeline uses a non-linear rectification scheme (Pollefeys et al., 1999b) that can deal with all types of camera motion (including forward motion). For the integration of multiple depth maps into a single surface representation a volumetric approach has been implemented (Curless and Levoy, 1996). The texture is obtained by blending the original images based on the surface geometry so that the texture quality is optimized. The resulting system is much more robust and accurate. This makes it possible to efficiently use it for many different applications.

2 FROM VIDEO TO 3D MODELS

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 corre-

* corresponding author

sponding image features between these images (i.e. image points that originate from the same 3D feature). In the case of video data features are tracked until disparities become sufficiently large so that an accurate estimation of the epipolar geometry becomes possible.

The next step consists of recovering the motion and calibration of the camera and the 3D structure of the tracked or matched 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 a priori calibration. Using a self-calibration algorithm (Pollefeys et al., 1999a) 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 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., 1999b). 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

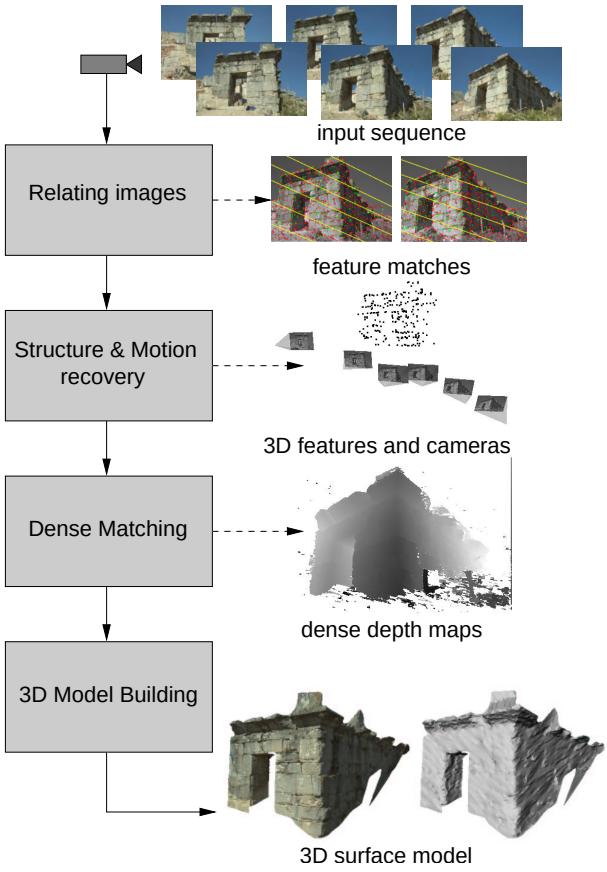


Figure 1: Overview of our image-based 3D recording approach.

matching or tracking (e.g. a pixel in a homogeneous region), feature points need to be selected (Harris and Stephens, 1988, Shi and Tomasi, 1994). Depending on the type of image data (i.e. video or still pictures) the feature points are tracked or matched 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 and Bolles, 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.

In case of video computing the epipolar geometry between two consecutive views is not well determined. In fact as long as the camera has not sufficiently moved, the motion of the features can just as well be explained by a homography. The Geometric Robust Information Criterion (GRIC) proposed by Torr (Torr et al., 1999) allows to evaluate which of the two models –epipolar geometry (F) or homography (H)– is best suited to explain the data. Typically, for very small baselines the homography model is always selected, as the baseline gets larger both models become equivalent and eventually the epipolar geometry model outperforms the homography based one. One can reliably compute the epipolar geometry from the moment

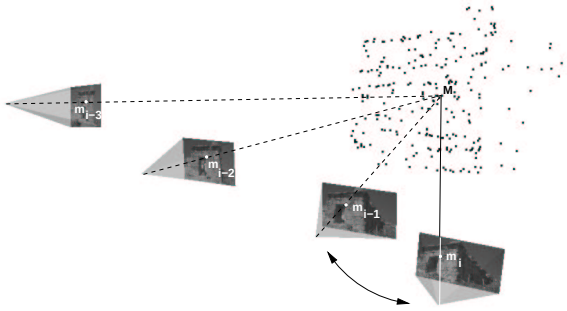


Figure 2: The pose estimation of a new view uses inferred structure-to-image matches.

that the F-GRIC value drops below the H-GRIC value. We then select as a new key-frame the last frame for which the number of tracked features is above 90% of the number of features tracked at the F-GRIC/H-GRIC intersection.

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) have been developed for this. Then the ambiguity is restricted to metric through self-calibration (Pollefeys et al., 1999a). 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.

If in some views all tracked feature are located on a plane, the approach explained above would fail. This problem can be detected and solved by using the approach proposed in (Pollefeys et al., 2002). A statistical information criterion is used to detect the images that only observe planar features and for these views the pose of the camera is only computed after the intrinsic camera parameters have been obtained through self-calibration (assuming they are all kept constant). In this way problems of ambiguities are avoided.

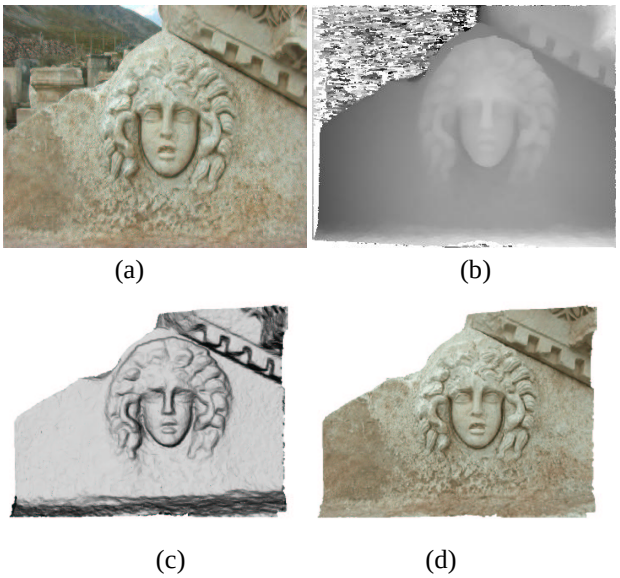
Figure 3: Example of a rectified stereo pair.

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 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., 1999b) 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., 1999b) 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



(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 and (d) textured view of the 3D model.

increased through the combination of multiple viewpoints and large global baseline while the matching is simplified through the small local baselines.

2.4 Building visual models

In the previous sections a dense structure and motion recovery approach was explained. This yields all the necessary information to build photo-realistic virtual models.

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, Ofek et al., 1997).

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 in (Curless and Levoy, 1996) is used.

An important advantage of our approach compared to more interactive techniques (Debevec et al., 1996, PhotoModeler) 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. A disadvantage of our approach (and more in general of most image-based approaches) is that our technique can not directly capture the photometric properties of an object, but only the combination of these with lighting. It is therefore not possible to re-render the 3D model under different lighting. This is a topic of future research.

lightfield rendering Alternatively, when the purpose is to render new views from similar viewpoints image-based approaches can be used (Levoy and Hanrahan, 1996, Gortler et al., 1996). The approach we present here avoids the difficult problem of obtaining a consistent 3D model by using view-dependent texture and geometry. This also allows to take more complex visual effects such as reflections and highlights into account. This approach renders views directly from the calibrated sequence of recorded images with use of local depth maps. The original images are directly mapped onto one or more planes viewed by a virtual camera.

To obtain a high-quality image-based scene representation, we need many views from a scene from many directions. For this, we can record an extended image sequence moving the camera in a zigzag like manner. To obtain a good quality structure-and-motion estimation from this type of sequence and reduce error accumulation it can be important to also match close views that are not predecessors or successors in the image stream (Koch et al., 1999).

The simplest approach consists of approximating the scene geometry by a single plane. The mapping from a recorded image to a new view or vice-versa then corresponds to a homography. To construct a specific view it is best to interpolate between neighboring views. The color value for a particular pixel can thus best be obtained from those views whose projection center is close to the viewing ray of this pixel or, equivalently, project closest to the specified pixel. For simplicity the support is restricted to the nearest three cameras (see Figure 5). All camera centers are projected into the virtual image and a 2D triangulation is performed. The cameras corresponding to the corners of a triangle then contribute to all pixels inside the triangle. The color values are blended using the baricentric coordinates on the triangle as weights. The total image is built up as a mosaic of these triangles. Although this technique assumes a very sparse approximation of geometry, the rendering results show only small ghosting artifacts (see experiments).

The results can be further improved. It is possible to use a different approximating plane for each triangle. This improves the accuracy further as the approximation is not done for the whole scene but just for that part of the image which is seen through the actual triangle. The 3D position of the triangle vertices can be obtained by looking up the

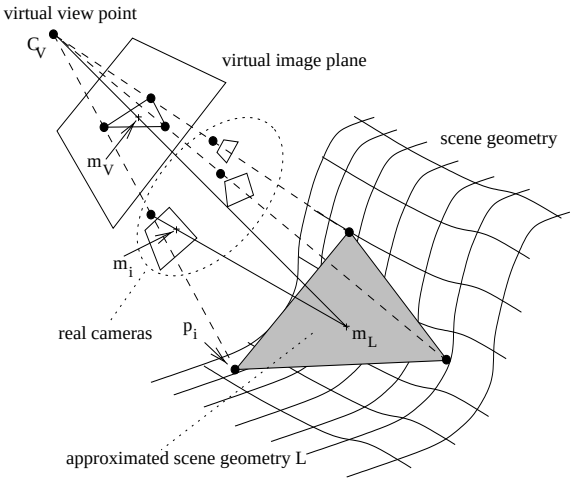


Figure 5: Drawing triangles of neighboring projected camera centers and approximating geometry by one plane for the whole scene, for one camera triple or by several planes for one camera triple.

depth value for the projection of the virtual viewpoint in the depth map correspondiong to each vertex. These points can be interpreted as the intersections of the lines connecting the virtual viewpoint and the real viewpoints with the scene geometry. Knowing the 3D coordinates of triangle corners, we can define a plane through them and apply the same rendering technique as described above.

Finally, if the triangles exceed a given size, they can be subdivided into four sub-triangles. For each of these sub-triangles, a separate approximative plane is calculated in the above manner. Of course, further subdivision can be done in the same way to improve accuracy. Especially, if just a few triangles contribute to a single virtual view, this subdivision is generally necessary. It should be done in a resolution according to performance demands and to the complexity of the geometry. Rendering can be performed in real-time using alpha blending and texture mapping facilities of todays graphics hardware. More details on this approach can be found in (Koch et al., 1999, Heigl et al., 1999, Koch et al., 2001). A similar approach was presented recently (Buehler et al., 2001).

We have tested our approaches with an image sequence of 187 images showing an office scene. Figure 6 (top-left) shows one particular image. A digital consumer video camera (Sony TRV-900) was swept freely over a cluttered scene on a desk, covering a viewing surface of about $1m^2$. Figure 6 (top-right) shows the calibration result. Result of a rendered view are shown in the middle of the figure. The image on the left is rendered with a planar approximation while the image on the right was generated with two levels of subdivision. Note that some ghosting artefacts are visible for the planar approximation, but not for the more detailed approximation. It is also interesting to note that most ghosting occures in the vertical direction because the inter-camera distance is much larger in this direction. In the lower part of Figure 6 a detail of a view is shown for the different methods. In the case of one global plane (left image), the reconstruction is sharp where the approximating plane intersects the actual scene geometry. The reconstruction is blurred where the scene geometry diverges from this plane. In the case of local planes (middle image), at the corners of the triangles the reconstruction is almost sharp, because there the scene geometry is considered directly. Within a triangle, ghosting artifacts occur where the scene geometry diverges from the particular local plane. If these triangles are subdivided (right image) these artifacts are reduced further.

Figure 6: Unstructured lightfield rendering: image from the original sequence (top-left), recovered structure and motion (top-right), novel views generated for planar (bottom-left) and view-dependent (bottom-right) geometric approximation.

3 CONCLUSION

In this paper an automatic approach was presented that takes a video sequence as input and computes a 3D model as output. By combining state-of-the-art approaches developed in the field of computer vision, computer graphics and photogrammetry, our system is able to obtain good quality results on video as well as on photographic material.

ACKNOWLEDGEMENTS

The authors are grateful to Marc Waelkens and his team for making the archaeological material accessible to them. Part of this work was carried out in collaboration with Reinhard Koch and Benno Heigl. The financial support of the FWO project G.0223.01, the IST projects INVIEW, ATTEST and 3DMurale are also gratefully acknowlegded. Kurt Cornelis is a research assistant of the Fund for Scientific Research - Flanders (Belgium).

REFERENCES

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

Buehler C, Bosse M, McMillan L, Gortler S, Cohen M. Unstructured Lumigraph Rendering, *Proc. SIGGRAPH* 2001, pp. 425-432.

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

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

O. Faugeras, 1992. 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.

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

Gortler S, Grzeszczuk R, Szeliski R and Cohen MF, The Lumigraph, *Proc. SIGGRAPH '96*, pp 43–54.

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

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

Heigl B, Koch R, Pollefeys M, Denzler J, Van Gool L. Plenoptic Modeling and Rendering from Image Sequences taken by Hand-held Camera, *Proc. DAGM'99*, pp. 94–101.

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

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

Koch, R., Heigl, B., Pollefeys, M., 2001. Image-Based Rendering from Uncalibrated Lightfields with Scalable Geometry, In R. Klette, T. Huang, G. Gimel'farb (Eds.), *Multi-Image Analysis*, Lecture Notes in Computer Science, Vol. 2032, pp.51-66, Springer-Verlag.

Levoy M, Hanrahan P, Lightfield Rendering, *Proc. SIGGRAPH '96*, pp 31–42.

Matas, J. and Chum, O., 2001. Randomized RANSAC, Center for Machine Perception, Czech Technical University, Prague, Research Report, CTU–CMP–2001–34, ftp://cmp.felk.cvut.cz/pub/cmp/articles/matas/matas-tr-2001-34.ps.gz (accessed 15 june 2002)

Ofek, E., Shilat, E., Rappopport, A., Werman, M., Highlight and Reflection Independent Multiresolution Textures from Image Sequences, *IEEE Computer Graphics and Applications*, 17(2).

PhotoModeler, Eos Systems Inc., http://www.photomodeler.com (accessed 15 June 2002).

Pollefeys, M., Koch, R., Van Gool, L., 1998. Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters, *Proc. International Conference on Computer Vision*, Narosa Publishing House, pp.90-95.

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

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

Pollefeys, M., Koch, R., Vergauwen, M., Van Gool, L., 2000. Automated reconstruction of 3D scenes from sequences of images, Isprs Journal Of Photogrammetry And Remote Sensing 55(4), pp. 251-267.

Pollefeys, M., Verbiest, F., Van Gool, L., 2002. Surviving Dominant Planes in Uncalibrated Structure and Motion Recovery, *Proc. ECCV*, Lecture Notes in Computer Science, Vol.2351, pp. 837-851.

Shi J., Tomasi, C., 1994. Good Features to Track, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pp. 593 - 600.

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

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

Torr, P., Fitzgibbon, A., Zisserman, A., 1999. The problem of degeneracy in structure and motion recovery from uncalibrated image sequences, *International Journal of Computer Vision*, 32(1) pp. 27-44.

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

Van Meerbergen, G., Vergauwen, M., Pollefeys, M., Van Gool, L., 2002. A Hierarchical Symmetric Stereo Algorithm Using Dynamic Programming, *International Journal of Computer Vision*, 47(1-3), pp. 275-285, 2002.