Virtual Models from Video and Vice-Versa

Marc Pollefeys, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Luc Van Gool

Center for Processing of Speech and Images, K.U.Leuven, Belgium

Summary. In this paper an approach is presented that obtains virtual models from sequences of images. The system can deal with uncalibrated image sequences acquired with a hand-held camera. Based on tracked or matched features the relations between multiple views are computed. From this both the structure of the scene and the motion of the camera are retrieved. The ambiguity on the reconstruction is restricted from projective to metric through auto-calibration. A flexible multi-view stereo matching scheme is used to obtain a dense estimation of the surface geometry. From the computed data virtual models can be constructed or, inversely, virtual models can be included in the original images.

1 Introduction

There has recently been a lot of interest in obtaining virtual models of existing scenes. Image-based approaches have shown a lot of potential in many areas. One of the areas where interesting applications exist is architecture. Nowadays most buildings are being designed on computer using CAD and visualization tools allow virtual visits. This can be very effective in presenting plans to persons that are not trained in reading them. However, most constructions have to be considered in their environment. It is therefore necessary to be able to generate a realistic impression of the environment too. Due to the complexity of natural sites a manual reconstruction can often not be considered and there is a need for more automated approaches that can directly capture the environment. Other applications can be found in the field of conservation of built heritage. In this area photogrammetric techniques have been used for many years. However, through advances in automation and digital technology much more complete analyses can be achieved at reduced cost. In addition, digital 3D models can also be used for planning restorations and as archives afterwards. Of course, there is also an important demand for photo-realistic models of monuments and sites for multi-media and entertainment products.

For most of the above 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. Some approaches have been proposed for extracting 3D shape and texture from image sequences acquired with a freely moving camera have been proposed. The approach of Tomasi and Kanade [20] used an affine factorization method to extract 3D from image sequences. An important restriction of this system is the assumption of orthographic projection. Another type of approach starts from an approximate 3D model and camera poses and refines the model based on images (e.g. *Façade* proposed by Debevec et al. [5]). The advantage is that less images are required. On the other hand a preliminary model must be available and the geometry should not be too complex.

The approach presented in this paper avoids most of these restrictions. The approach 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.

2 Relating images

Starting from a collection of images or a video sequence the first step consists in relating the different images to each other. This is not a 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), the first step consist of selecting feature points [10,19]. These are suited for tracking or matching. 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 be contaminated with an important number of wrong matches or *outliers*. In this case, a traditional least-squares approach will fail and therefore a robust method is used [21,9]. Once the multi-view constraints have been obtained they can be used to guide the search for additional correspondences. These can then be used to further refine the results for the multi-view constraints.

3 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. The approach that is used is related to [1] but is fully projective and therefore not dependent on the quasi-euclidean initialization. This is achieved by strictly carrying out all measurements in the images, i.e. using reprojection errors instead of 3D errors. To support initialization and determination of close views (independently of the actual projective frame) an image-based measure to obtain a qualitative evaluation of the distance between two views had to be used. The proposed measure is the minimum median residual for a homography between the two views.

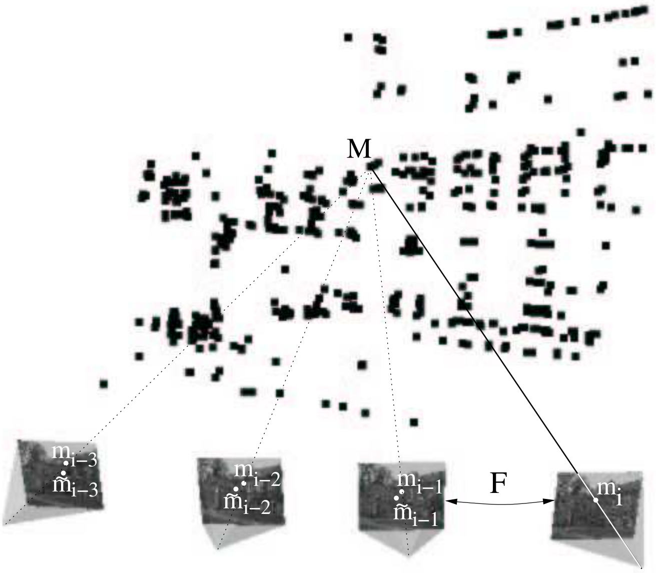

Fig. 1. The pose estimation of a new view uses inferred structure-to-image matches.

At first two images are selected and an initial projective reconstruction frame is set-up [7,11]. Then the pose of the camera for the other views is determined in this frame and for each additional view the initial reconstruction is refined and extended. This is illustrated in Figure 1. In this way the pose estimation of views that have no common features with the reference views also becomes possible. Typically, a view is only matched with its predecessor in the sequence. In most cases this works fine, but in some cases (e.g. when the camera moves back and forth) it can be interesting to also relate a new view to a number of additional views [15]. Candidate views are identified using the image-based measure mentioned above. Once the structure and motion has been determined for the whole sequence, the results can be refined through a projective bundle adjustment [23]. Then the ambiguity is restricted to metric through auto-calibration [8]. Our approach is based on the concept of the absolute quadric [22,18]. Finally, a metric bundle adjustment is carried out to obtain an optimal estimation of the structure and motion.

4 Dense surface estimation

To obtain a more detailed model of the observed surface dense matching 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 [17] is used. This approach can deal with arbitrary relative camera motion and guarantees minimal image sizes while standard homography-based approaches fail when the epipole is contained in the image. 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 2. It was recorded with a hand-held digital video camera in the Béguinage in Leuven. Due to the narrow streets only forward motion is feasible. This would have caused standard rectification appraoches to fail.

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 . 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. Matching ambiguities are resolved by exploiting the ordering constraint in the dynamic programming approach [13]. 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 [6]. The disparity search range is limited based on the disparities that were observed for the features in the structure and motion recovery.

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. This approach was discussed more in detail in [14].

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.

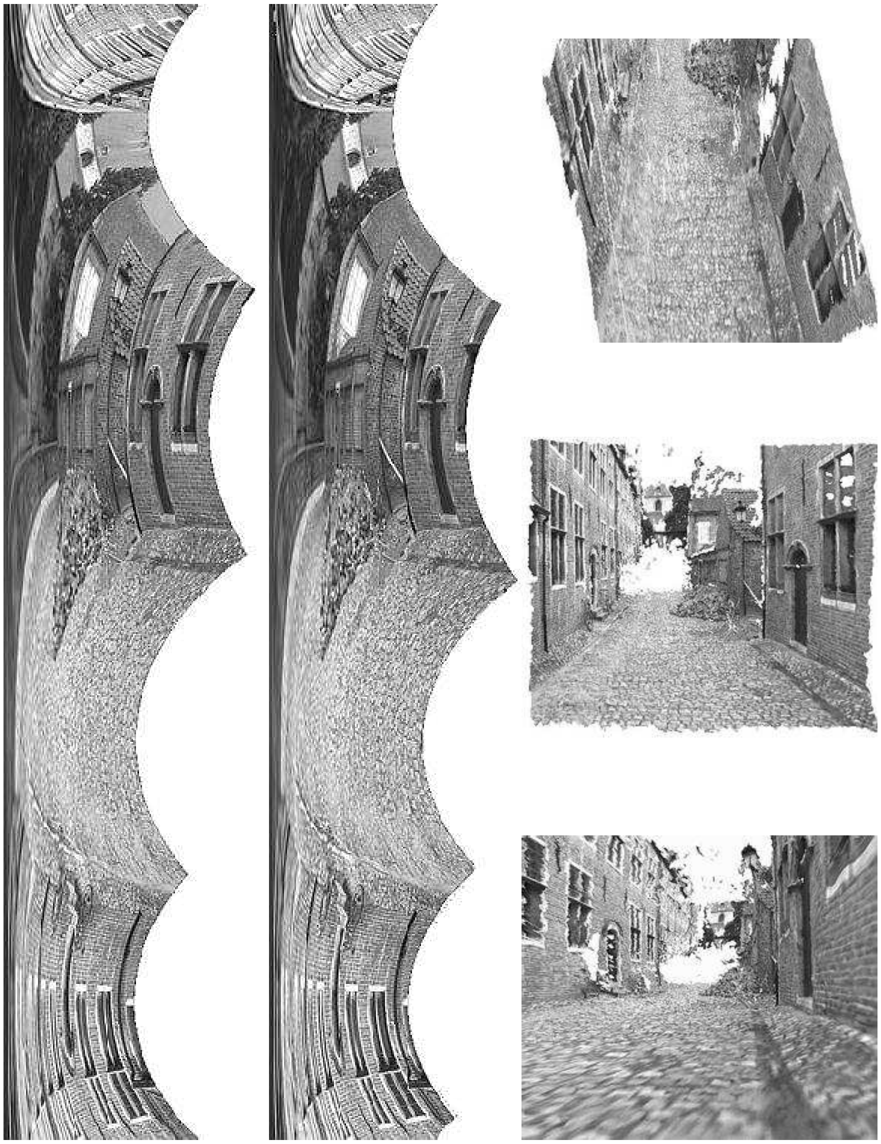

Fig. 2. *Béguinage* sequence: Rectified image pair (left) and some views of the reconstructed street model obtained from several image pairs (right).

5 Building virtual models

In the previous sections a dense structure and motion recovery approach was given. This yields all the necessary information to build photo-realistic virtual 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 and is illustrated in Figure 3. Some more views can also be seen in Plate 1 Fig. 1.

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[16].

To reconstruct more complex shapes it is necessary to combine multiple depth maps. Since all depth-maps can be located in a single metric frame, registration is not an issue. In some cases it can be sufficient to load the separate models together in the graphics system. For more complex scenes it can be interesting to first integrate the different meshes into a single mesh. This can for example be done using the volumetric technique proposed in [4].

Alternatively, when the purpose is to render new views from similar viewpoints image-based approaches can be used [15,12]. This approach 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.

The *Indian temple* sequence was shot in Ranakpur (India) using a standard Nikon F50 photo camera and then scanned. The sequence seen at the top of Figure 4 was processed through the method presented in this paper. The results can be seen in the middle and lower part of Figure 4. Some more detailed views can be seen in Figure 5. Note that some of these artificial views are taken under viewing angles that are very different from the original pictures. This shows that the recovered models allow to extrapolate viewpoints to some extent.

6 Fusion of real and virtual scenes

Another interesting possibility offered by the presented approach is to combine real and virtual scene elements. This allows to augment real environments with virtual objects or vice-versa. A first approach consists of virtualizing the real environment and then to place virtual objects in it. The

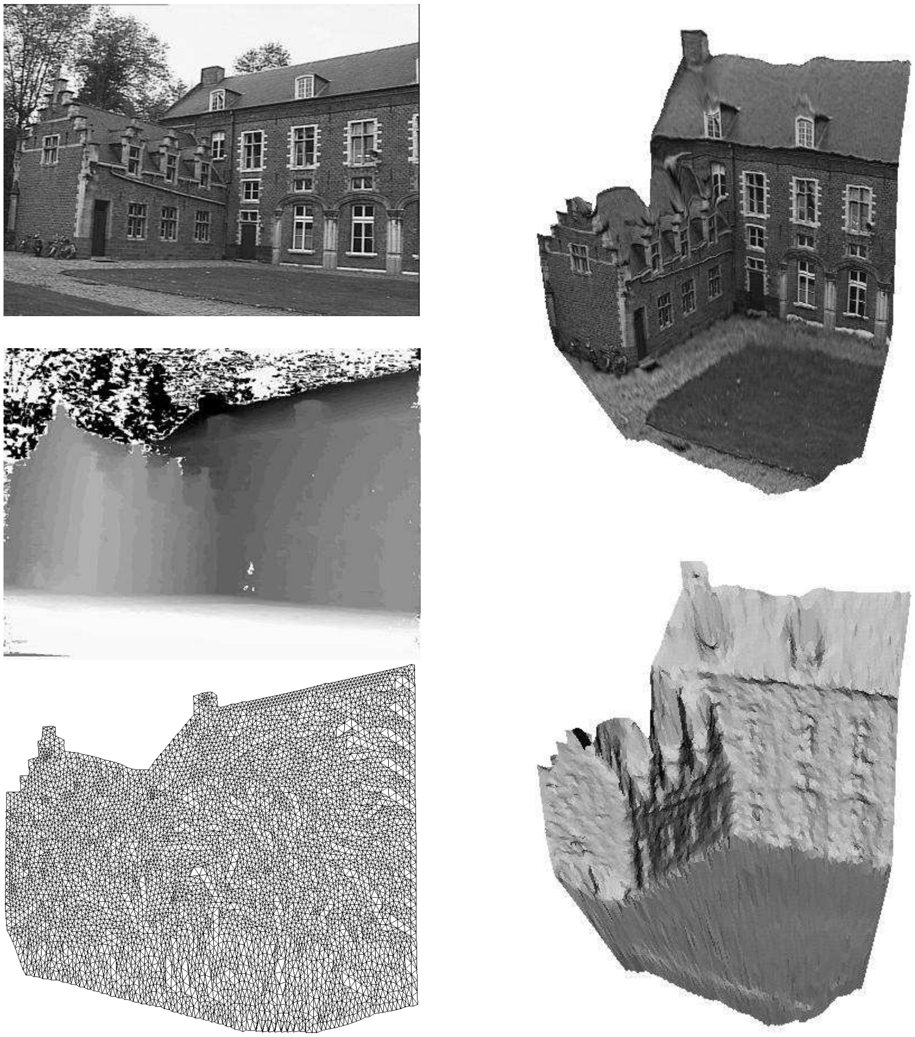

Fig. 3. Surface reconstruction approach (left): A triangular mesh is overlaid on top of the image. The vertices are back-projected in space according to the depth values. From this a 3D surface model is obtained (right)

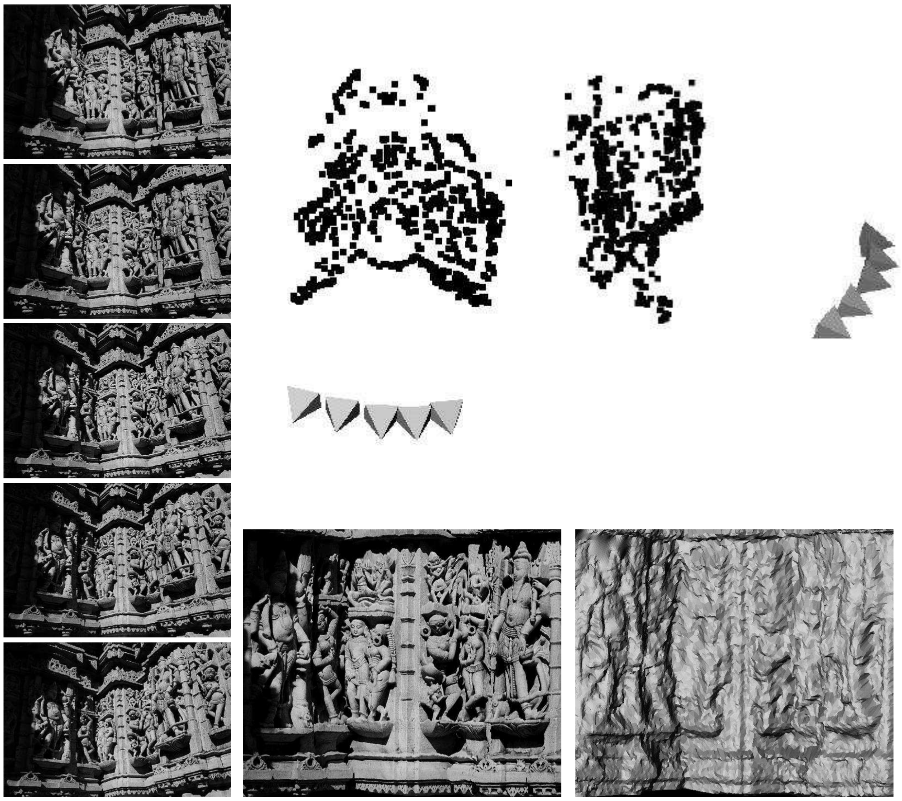

Fig. 4. The *Indian temple* sequence (left), recovered sparse structure and motion (top-right) and textured and a shaded view of the reconstructed 3D surface model (bottom-right).

landscape of Sagalassos (an archaeological site in Turkey) was modeled from a dozen photographs taken from a nearby hill. Virtual reconstructions of ancient monuments have been made based on measurements and hypotheses of archaeologists. Both could then be combined in a single virtual world. A view is shown in Plate 1 Fig. 1 (middle).

Another challenging application consists of seamlessly merging virtual objects with 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 are the rigid registration of virtual objects into the real environment, the problem of mutual occlusion of real and virtual objects and the extraction of the illumination distribution of the real environment in order to render the virtual objects with this illumination model.

Here we will concentrate on the first of these problems, although the computations described in the previous section also provide most of the necessary

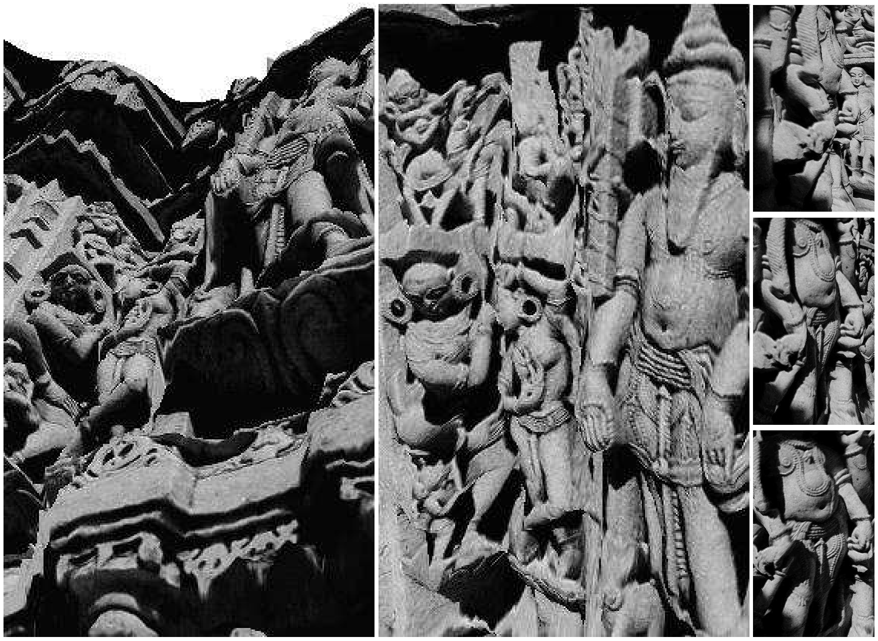

Fig. 5. Some more detailed views of the *Indian temple* reconstruction.

information to solve for occlusions and other interactions between the real and virtual components of the augmented scene. Accurate registration of virtual objects into a real environment is still a challenging problems. Systems that fail to do so will also fail to give the user a real-life impression of the augmented outcome. Since our approach does not use markers or a-priori knowledge of the scene or the camera, this allows for us to deal with video footage of unprepared environments or archive video footage. More details on this approach can be found in [2].

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 features should be tracked from frame to frame. A key component in this case is the bundle adjustment. It does not only reduce the frame to frame jitter, but removes the largest part of the error that the structure and motion approach accumulates over the sequence. According to our experience it is very important to extend the perspective camera model with at least one parameter for radial distortion to obtain an undistorted metric structure (this will be clearly demonstrated in the example). Undistorted models are required to position larger virtual entities correctly in the model and to avoid drift of virtual objects in the augmented video sequences.

The following example was recorded at Sagalassos in Turkey, where footage of the ruins of an ancient fountain was taken. The *fountain* video sequence consists of 250 frames. 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. The top part of Figure 6 shows a top view of the recovered structure before and after bundle-adjustment. Besides the larger reconstruction error it can also be noticed that the non-refined structure is slightly bent. This effect mostly comes from not taking the radial distortion into account in the initial structure recovery. In the rest of Figure 6 some frames of the augmented video are shown. Two frames are also shown in Plate 1 Fig. 1 (bottom).

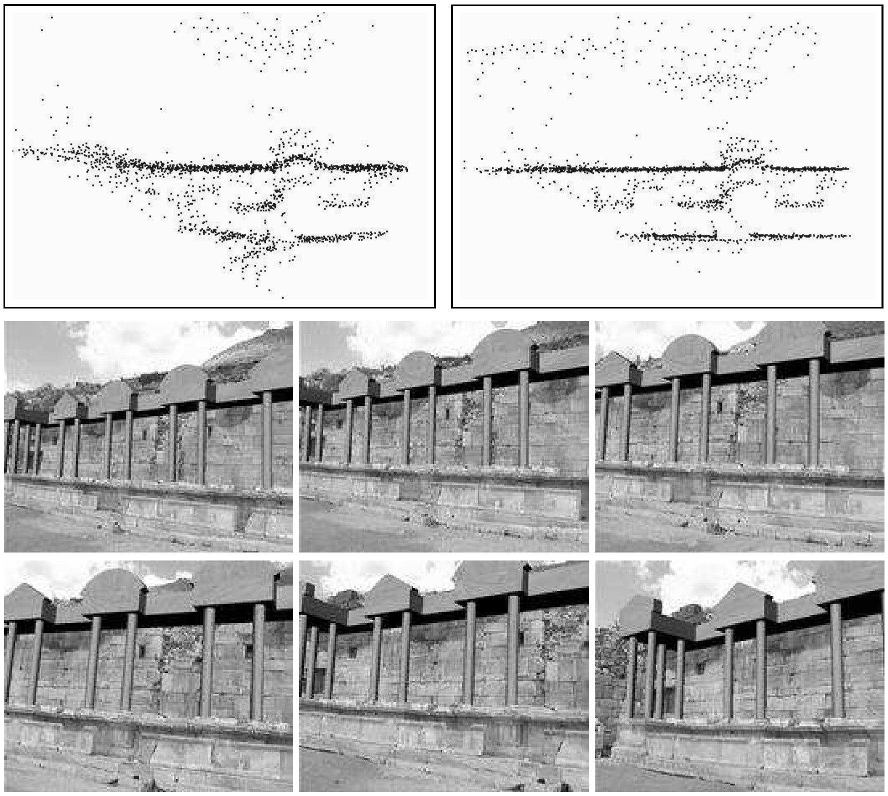

Fig. 6. Fusion of real and virtual fountain parts. Top: recovered structure before and after bundle adjustment. Bottom: 6 of the 250 frames of the fused video sequence

7 Conclusion

In this paper an approach for obtaining virtual models with a hand-held camera 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. From these results virtual models can be obtained or, inversely, virtual models can be inserted in the original video.

Acknowledgement

Marc Pollefeys and Kurt Cornelis are respectively post-doctoral fellow and research assistant of the Fund for Scientific Research - Flanders (Belgium). The financial support of the FWO project G.0223.01, the ITEA BEYOND and the STWW VirtErf projects of the IWT and the IST-1999-26001 project VIBES are also gratefully acknowlegded.

References

1. P. Beardsley, A. Zisserman and D. Murray, "Sequential Updating of Projective and Affine Structure from Motion", *International Journal of Computer Vision* (23), No. 3, Jun-Jul 1997, pp. 235-259.
2. 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.
3. Cox, I., Hingorani, S., Rao, S., 1996, A Maximum Likelihood Stereo Algorithm, Computer Vision and Image Understanding, Vol. 63, No. 3.
4. B. Curless and M. Levoy, "A Volumetric Method for Building Complex Models from Range Images" *Proc. SIGGRAPH '96*, pp. 303–312, 1996.
5. 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.
6. Falkenhagen, L., 1997, Hierarchical Block-Based Disparity Estimation Considering Neighbourhood Constraints. Proceedings International Workshop on SNHC and 3D Imaging, Rhodes, Greece, pp.115-122.
7. 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.

8. O. Faugeras, Q.-T. Luong and S. Maybank. “Camera self-calibration: Theory and experiments”, *Computer Vision - ECCV’92*, Lecture Notes in Computer Science, Vol. 588, Springer-Verlag, pp. 321-334, 1992.
9. 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.
10. C. Harris and M. Stephens, “A combined corner and edge detector”, *Fourth Alvey Vision Conference*, pp.147-151, 1988.
11. R. Hartley, R. Gupta, and T. Chang. “Stereo from uncalibrated cameras”. Proc. Conference Computer Vision and Pattern Recognition, pp. 761-764, 1992.
12. B. Heigl, R. Koch, M. Pollefeys, J. Denzler and L. Van Gool, Plenoptic Modeling and Rendering from Image Sequences taken by Hand-held Camera, Proc. DAGM’99, pp.94-101.
13. Koch, R., 1996, Automatische Oberflachenmodellierung starrer dreidimensionaler Objekte aus stereoskopischen Rundum-Ansichten, PhD thesis, University of Hannover, Germany, also published as Fortschritte-Berichte VDI, Reihe 10, Nr.499, VDI Verlag, 1997.
14. 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.
15. R. Koch, M. Pollefeys, B. Heigl, L. Van Gool and H. Niemann. “Calibration of Hand-held Camera Sequences for Plenoptic Modeling”, *Proc.ICCV’99 (international Conference on Computer Vision)*, pp.585-591, Corfu (Greece), 1999.
16. E. Ofek, E. Shilat, A. Rappopport and M. Werman, “Highlight and Reflection Independent Multiresolution Textures from Image Sequences”, *IEEE Computer Graphics and Applications*, vol.17 (2), March-April 1997.
17. 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.
18. 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.
19. J. Shi and C. Tomasi, “Good Features to Track”, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR’94)*, pp. 593 - 600, 1994.
20. C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: A factorization approach”, *International Journal of Computer Vision*, 9(2):137-154, 1992.
21. P. Torr, *Motion Segmentation and Outlier Detection*, PhD Thesis, Dept. of Engineering Science, University of Oxford, 1995.
22. B. Triggs, “The Absolute Quadric”, *Proc. 1997 Conference on Computer Vision and Pattern Recognition*, IEEE Computer Soc. Press, pp. 609-614, 1997.
23. 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.

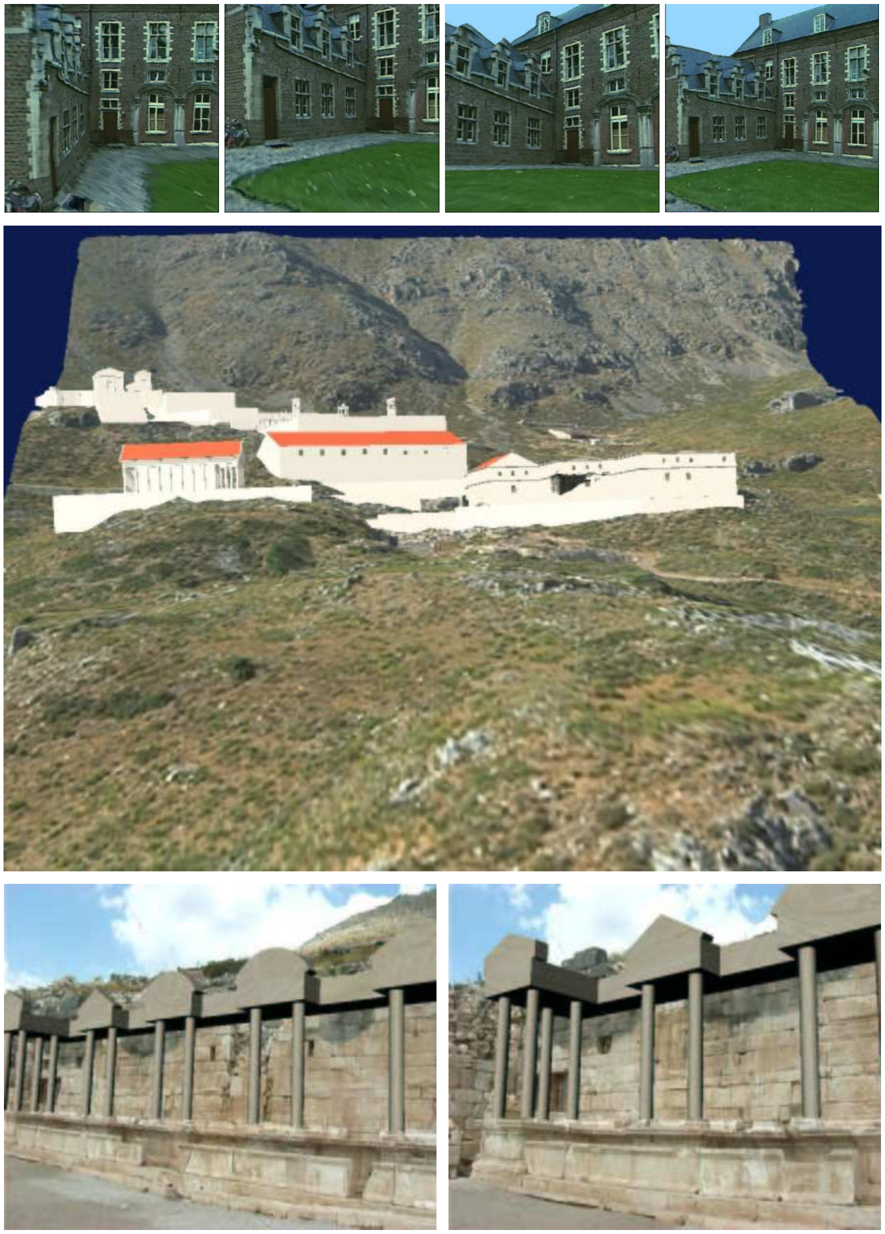

Plate 1 Fig. 1 Some views of the reconstructed *castle* (top), Virtualized landscape of Sagalassos combined with virtual reconstructions of monuments (middle), Two frames of augmented *fountain* video sequence.