VR Modeler: From Image Sequences to 3D Models

Mario Sormann * Joachim Bauer Christopher Zach Andreas Klaus
Konrad Karner

VRVis Research Center

Abstract

In this paper we present a novel interactive modeling system, called *VR Modeler*, to create 3D geometric models from a set of photographs. In our approach standard automatic reconstruction techniques are assisted by a human operator. The modeling system is efficient and easy to use because the user can concentrate on the 2D segmentation and interpretation of the scene whereas our system is responsible for the corresponding 3D information. Therefore we developed the user interface of VR Modeler as a monocular 3D modeling system. Additionally, we are able to obtain coarse as well as high resolution models from architectural scenes. Finally, we tested the modeling system on different types of datasets to demonstrate the usability of our approach.

CR Categories: I.4.5 [Image Processing and Computer Vision]: Reconstruction—Transform Methods; I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Modeling Packages

Keywords: interactive modeling system, user interface, feature based modeling, photogrammetry, 3D reconstruction, image sequences, 3D modeling

1 Introduction

The creation of 3D models for use in an interactive virtual environment is an expensive and tedious process and is still a challenging problem in computer vision. Typically the requirement that the virtual environment should mirror an existing scene demands accurate three dimensional (3D) geometry, as well as surface materials or textures. Thus, there is a need for a method to directly extract realistic 3D models from real photographs.

In the fields of photogrammetry and computer vision many approaches have been developed which allow the production of photorealistic 3D models [Pollefeys et al. 2000], [Zisserman et al. 2000]. In general, these algorithms take multiple images of a real environment using a calibrated camera and then create from these images a 3D structure of the scene. The output of such an algorithm is a dense point cloud, corresponding to important features in the scene. These point clouds should be converted into logical objects in order to create suitable representations for a virtual environment. Current available methods for automatic segmentation are not yet robust enough to build useful geometric models for the visualization, thus fully automatic segmentation yields to an ill-posed poblem.

In this paper we discuss how we can make the modeling process more convenient and efficient. So far there are two separate research areas in computer vision, one is the reconstruction problem and the other one the recognition problem. In our approach we solve the reconstruction problem by highly redundant information about the scene, in our case image sequences. The recognition problem is handed over to a human operator, who is supported by an intelligent user interface. Thus the operator can focus on the segmentation and interpretation of the scene using only one image while the system takes care about the associated 3D information.

Essentially our interactive modeling system, called *VR Modeler* (Virtual Reality Modeler) allows a user to construct a geometric model of the scene from a set of photographs. The images are taken with a hand-held digital consumer camera using short baselines. After some preprocessing the relative orientation of the image sequences are calculated fully automatic. Our orientation method, which is not the topic of this paper, is based on work described by Horn [Horn 1990], Klaus et al. [Klaus et al. 2002] and Nister [Nister 2003]. Once we have determined the relative orientation between all image pairs we are able to extract 3D information from the photographs automatically by employing area and feature based matching techniques. The 3D information consists of 3D points, 3D lines and 3D surfaces, as illustrated in Figure 1.

After applying this automatic reconstruction process all

*sormann@vrvis.at

Figure 1: Fusion of automatically extracted 3D lines and 3D point cloud. The illustration shows the bell tower on the castle hill in Graz.

2D features in the images correspond to their 3D counterparts. Due to the fact that we like to obtain a consistent 3D model of the scene it is necessary to combine the extracted different model representations. In our approach we decided to accomplish this task by utilizing a human operator in terms of his interpretation and segmentation abilities. As a result of the modeling process we are able to achieve a coarse as well as a detailed 3D model of the scene.

The remainder of this paper is structured as follows: after a section related work we describe our approach to build 3D geometric models from image sequences. Furthermore we present the achieved results and finally we conclude our approach and outline some aspects of future work.

2 Related Work

The process of reconstructing 3D models from image sequences is a very active research topic in computer vision. Nevertheless no general technique exists to obtain fully automatic 3D models from image sequences. However three different research fields provide methods to recover 3D information from oriented digital images. These research fields are known as area based modeling, feature based modeling and human assisted reconstruction from image information.

2.1 Area based Modeling

The estimation of dense 3D point clouds from image sequences is discussed by Pollefeys et al. [Pollefeys et al. 2000] and Brown et al. [Brown et al. 2003]. The geometrical theory of these methods relies on being able to solve the reconstruction problem. From corresponding points the relative orientation can be estimated and 3D points are extracted. These procedure is applied to many pixels within stereo or multiview images which results in dense 3D point clouds.

2.2 Feature Based Modeling

Several authors discussed the problem of 3D data acquisition from digital images using various feature extraction and matching methods. A general overview of these methods is given in Baillard et al. [Baillard et al. 1999], where they propose a line matching method over multiple oriented views. Schmid and Zisserman [Schmid and Zisserman 2000] assume a 2D feature extraction method from images including contour chains, line segments and vanishing points to automatically recover planes from architectural images.

2.3 Human assisted Reconstruction from Image Information

One of the most popular approaches in this field is the modeling and rendering of architecture proposed by Debevec [Debevec 1996], called *Facade*. This approach, which combines geometry-based modeling and image-based modeling, can be separated into two main components. The first component facilitates the recovery of a basic geometric model of the photographed scene. The second component describes an efficient view dependent texturing method to better represent geometric details of the basic model. Another project named *Realise* [Leymarie et al. 1996] is based on a hybrid approach with vision techniques to assist users to extract models from image sequences. A user specifies interactively the topology of the scene whereas the system reconstructs the geometry from the images. The commercial product Image Modeler from Realviz [Modeler 2004] is inspired by both approaches.

3 VR Modeler: 3D Models from a Image Sequences

In this section we present the underlying model representations and algorithms. Furthermore we discuss some

user interface aspects illustrated on the implementation of VR Modeler. As already mentioned the input images are captured with a digital consumer camera using short baselines, thus a digital video camera with a reasonably high resolution will work as well. The extracted 3D primitives (3D points, 3D lines) and 3D surfaces acquired with current state of the art techniques will be insufficient in terms of interpretation, segmentation and visualization aspects. To make this aspects more concrete we will outline various problems of modeling 3D scenes in more detail. Consequently in this paper we focus on three aspects:

- The segmentation and interpretation of a scene to obtain a consistent 3D model from image sequences
- How the modeling process is affected by the visualization aspect
- How the user interface can make the process of modeling more convenient and accurate.

Figure 2 shows the Herz-Jesu church in Graz represented as a 3D point cloud. Obviously it is very difficult to obtain fully automatic a correct segmentation and interpretation of the scene from this model representation. In fact the segmentation and interpretation task includes the localization and classification of facades, windows, doors or any other relevant scene objects. Therefore in our opinion the better choice to acquire 3D models is to combine 3D surface representations with feature based modeling assisted by a human operator.

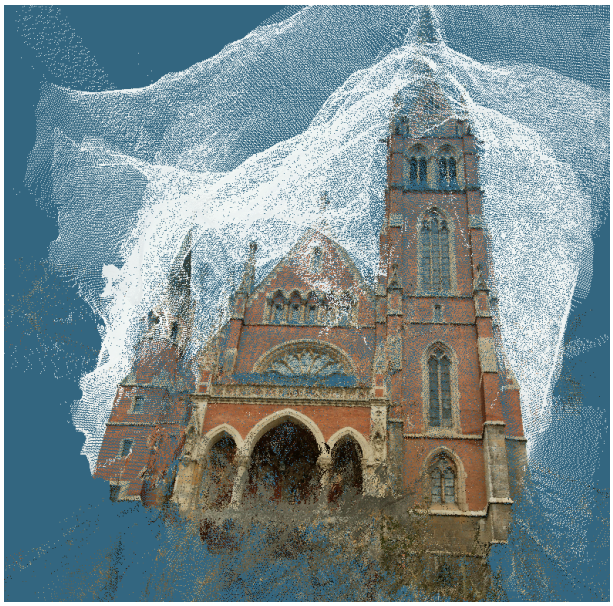

Figure 2: 3D point cloud from of the Herz-Jesu church in Graz. The generated 3D point cloud includes many outliers from areas like the sky, thus a segmentation and interpretation of the scene is needed.

The second question yields to the level of detail concept, where geometry, which is to complex to be rendered fast enough, is replaced by a simpler model. We observed that standard simplification methods are not well suited to generate different levels of detail for architectural models, since they do not preserve for example upright walls to which humans are very sensitive. Hence, we decided to produce a coarse as well as a high detailed polygonal models of the scene. Furthermore with such a coarse polygonal model we can achieve realtime rendering in high quality even on low bandwidth network connections [Zara and Slavik 2003].

The last question is related to the user interface of VR Modeler. In general user interfaces represent a key concept to computer applications and directly relate to the usability of a given application [Schneiderman 1998]. The key concept of our interactive modeling system is based on the fact that humans are not good at precise or accurate operations in 3D. Therefore we developed our VR Modeler as a so called monocular 3D modeling system.

3.1 3D Model Representations

In the VR Modeler, an architectural building is represented as a set of different model representations. Various representations can be found in [Bauer et al. 2002] and [Klaus et al. 2002]. Our approach deals with the following three types of model representations: marker points, marker lines and 3D point clouds. This ordering also reflects the complexity of the geometric primitives. The following sections give a detailed description of the mentioned representation types.

Marker Points

The extraction of marker points from image sequences is related to the reconstruction problem, which is to identify the 2D points in two images that are projections of the same 3D point in the world. From corresponding points within the image sequences the relative orientation and the 3D positions of the corresponding points can be estimated. Once we have determined the relative orientation, additional 3D points can be easily extracted from two 2D points in the image sequence, as illustrated in Figure 3. In VR Modeler we distinguish between a semi automatic and a fully automatic marker point extraction. The semi automatic extraction is supported over an incremental and straightforward process by a human operator. Consequently the user defines marker points in the image sequence over a simple point and click interface with subpixel accuracy by zooming into the images.

The automatic marker point generation is based on a standard point- -of-interest detector introduced by Harris and Stephens [Harris and Stephens 1988] followed by an automatic matching procedure [Brown et al. 2003]. The output of both procedures is a direct assignment of 2D marker points and their 3D counterpart.

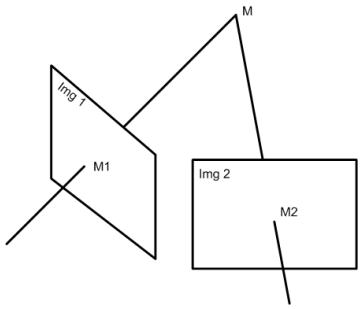

Figure 3: Reconstruction of a 3D point from two 2D points over a known relative orientation.

Marker Lines

Man-made objects, for example architectural buildings, require the usage of so called marker lines for the modeling step. Marker lines provide higher accuracy and are simpler to localize during the modeling step than marker points. The extraction of 2D line segments is based on the method proposed by Rothwell et al. [Rothwell et al. 1995]. Utilizing this algorithm it is possible to extract contour chains with subpixel accuracy. After applying a RANSAC [Fischler and Bolles 1981] based line detection method for all extracted contour chains and an optimization step based on vanishing points [Bauer et al. 2002], we derive a set of 2D line segments. These line segments in combination with the known relative orientation represent the input for our automatic line matching algorithm. The overall procedure of our 3D line matching method is described in [Schmid and Zisserman 2000]. Similar to marker points the outcome of the algorithm yields to a direct connection between 2D and 3D marker lines.

3D Point Cloud

The generation of dense 3D point clouds from calibrated images is performed by an iterative and hierarchical matching procedure exploiting the already known epipolar geometry between the images. For every sampling point the matching procedure optimizes a cost function, which contains the similarity between the template windows and a regularization term to favor smooth surfaces in textureless regions. Using a hierarchical approach

Figure 4: Automatically extracted 3D line set from the Herz-Jesu church in Graz.

many problems with repetitive patterns, as often encountered with building facades, are resolved [Brown et al. 2003]. Figure 5 shows a 3D point cloud of the clock tower on the castle hill in Graz.

Another possibility to incorporate 3D surfaces into the VR Modeler is obviously achievable with 3D laser scanning data.

3.2 Coarse 3D Models

The construction of coarse 3D models within VR Modeler is an incremental process supported by an human operator. Our approach allows the user to create 3D surfaces applying two different techniques.

The key concept of the first strategy is based on a selection of the interesting area of the scene by a human operator. In consideration of this segmentation issues a user connects in the image space the available marker points to a polygon, whereas the triangulation of the polygon and the 3D surface generation is performed by our system.

In a second method we emphasize marker lines to reconstruct polygons from the image sequence. Consequently the user supplies the input to the VR Modeler by specifying and grouping the marker lines in one of the images to facades, windows or doors. As a consequence of the direct link between the 2D and 3D model representation the 3D surface can be easily extracted.

Figure 5: 3D point cloud of the clock tower on the castle hill in Graz.

Texturing

Typically, texturing the reconstructed polygon from one image produce various disturbing artefacts, for instance occlusions yield to incorrect textured polygons. Therefore a multi view texturing approach [Bornik et al. 2002] allows to texture a polygon from all images more accurately and additionally increases the visual quality of the scene. Hence the texture information of the polygon is generated from all images in which the polygon is visible.

Results

Figure 6 shows a coarse model of the bell tower on the castle hill in Graz. Note, that this model has been created with the VR Modeler in 5 approximately minutes. This time takes into account all 3D modeling steps except the automatic orientation of the image sequence.

3.3 Detailed 3D Models

So far we have described the creation of a coarse model of the scene, but in general a facade of an architectural building will have additional geometric details which are not presented in the basic model. Therefore this section is dedicated to explain the creation of this detailed 3D models. As already outlined in section 3.1 we automatically recover a 3D point cloud from the image sequence of the scene. This point cloud and the marker points, respectively the marker lines are further used as an input for

(a) Front side

(b) Back side

Figure 6: Two views of the reconstructed bell tower on the castle hill in Graz.

our detailed modeling process. Note, that we obtain in both cases a segmentation and interpretation of the scene in meaningful units, like windows, doors or roofs. Additionally the the whole modeling procedure is supported by a human operator.

As outlined the segmentation process to create a coarse 3D model is based on two different techniques. We either utilize marker points or marker lines to obtain a segmented area of the scene. Obviously the detailed reconstruction is performed exclusively inside of this emphasized borderline. To obtain a polygonal 3D surface representation a standard image-based triangulation method can be finally performed.

Basically the algorithm works as follows: The already known region of interest provides a set of 2D line segments. Each of this 2D line segment corresponds to a 3D marker line, which is further utilized to construct an object plane. In the next step the final plane parameters are computed with a robust least-squares fit to the 3D lines endpoints. Finally, the acquired object plane is merged with the 3D point cloud, thus that we obtain a high resolution model of the segmented area of the scene. This

process is illustrated in Figure 7 where the red area indicates the acquired object plane.

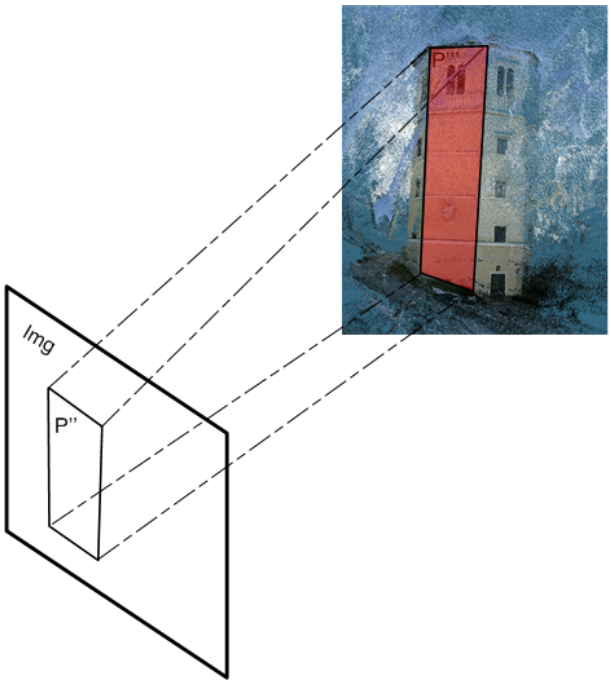

Figure 7: Illustration of the reconstruction process: All selected marker lines are used to compute a object plane which is merged with the previous extracted 3D point cloud. The red area indicates the acquired plane.

Detailed modeling of the scene via marker points or marker lines is advantageous for a number of reasons:

- Typical problems of area based modeling approaches are avoided, hence the final high resolution model is free of disturbing outliers
- As a consequence of the previous mentioned constraints a significant increase of the performance can be achieved
- Almost all architectural models can be easily created by arrangement of various polygons

The first point represents the main advantage of our interactive modeling system, which is shown in more detail in the following section.

Results

Figure 8 illustrates the raw 3D point cloud, whereas in Figure 9 the result of our detailed reconstruction approach is shown. Another high resolution model, which presents the bell tower on the castle hill in Graz is illustrated in Figure 10. As expected in both models the segmented regions are free of disturbing artifacts.

3.4 User Interface

The automatic creation of digital 3D models from images of real objects can be split into two main components: the user component, which is represented by simple human interaction and the computer component which is represented by more or less computational complex algorithms. In general user interfaces are directly related to the behavior of the human interaction. Therefore we designed the user interface as a monocular 3D modeling system, which emphasizes the advantages of 2D segmentation and interpretation.

Figure 11 illustrates our implemented user interface which contains two types of windows: image viewer and model viewer. Typically the user supplies the input to the program by utilizing the image viewer, whereas the model viewer is used to verify the reconstruction progress. The bottom image preview box comprises an overview of the captured photographs, thus a human operator can easily select the appropriate image for the reconstruction process.

Since the reconstruction problem is already solved a human operator can concentrate on the segmentation and interpretation of the scene. Therefore the general idea of the user interface is based on the fact that humans are clumsy at simultaneously controlling multiple degrees of freedom. Furthermore they are not good at precise or accurate operations in 3D, especially with a 2D interface such as a standard monitor and mouse. In contrast to humans computers are the better 3D operators, because they are not limited to two eyes. Additionally, they are able to handle multiple views simultaneously.

Due to this facts we implemented our user interface as a monocular 3D modeling system, where the user is responsible for the segmentation and interpretation in 2D, while the modeling system deals with the corresponding 3D information. Another benefit of this concept is that we obtain a full interpretation of the scene in logical units, like windows, roofs, doors, facades etc. These are the main differences between our user interface and those proposed in [Debevec 1996] and [Leymarie et al. 1996].

4 Conclusion and Future Work

We have presented a method to semi-automatically reconstruct virtual environments from a set of photographs. Furthermore we have discussed the underlying model representations, as well as our incremental method for 3D model building. Consequently our approach can be separated into two main components. The first component is an convenient interactive modeling system to recover a coarse geometric model of the scene. The second

Figure 8: Illustration of the Herz-Jesu church as a raw 3D point cloud.

Figure 9: Different views of the reconstruction result. In contrast to the previous illustration the reconstructed regions are free of disturbing artifacts.

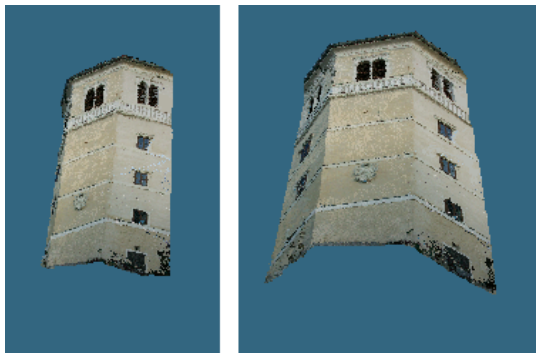

Figure 10: Two views of the front side of the bell tower represented as high resolution model.

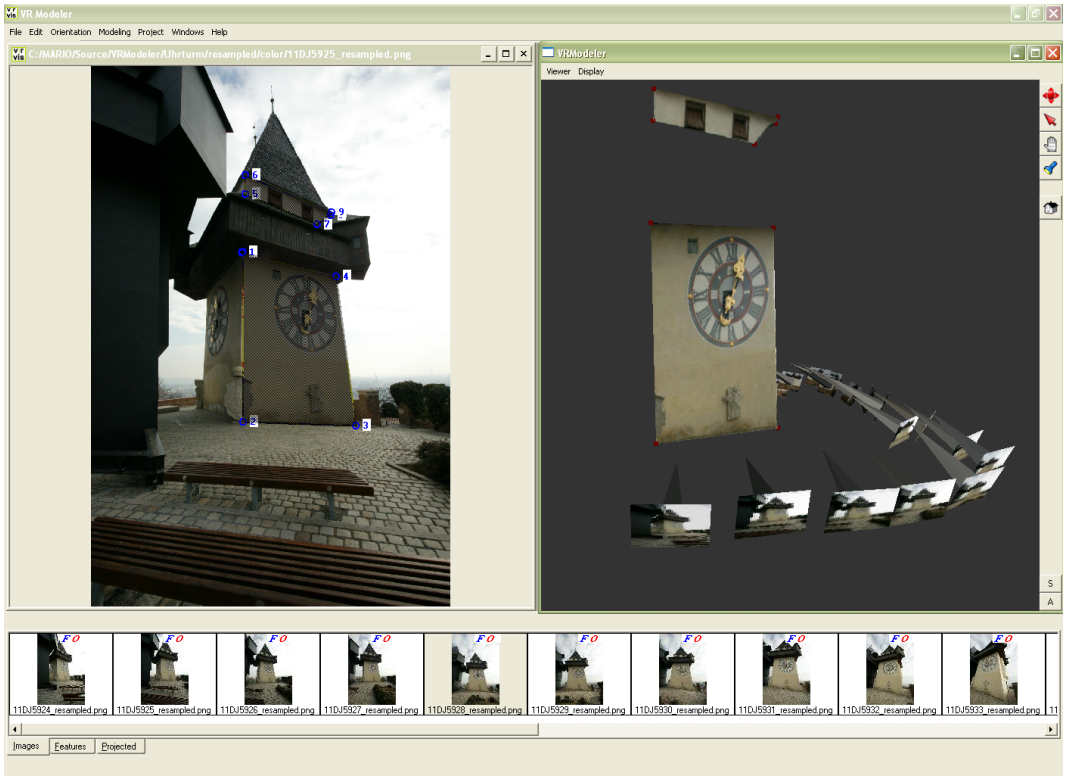

Figure 11: Overview of the user interface of VR Modeler, implemented as a monocular 3D modeling system.

component represents a reconstruction system to generate an accurate high resolution model of the scene.

Additionally, in *VR Modeler* we focus on the segmentation and interpretation of the scene, which is supported by an intelligent user interface. During the modeling process the need for manual interaction should be minimized to obtain a nearly automatic reconstruction. While the results are very promising and already satisfying for many scenes, improvements both in the modeling as well as in the user interface are suggested.

Future work includes evaluating the accuracy of geometric reconstructions and improve the functionality of the user interface. Further we will concentrate on an almost automatic 3D reconstruction based on additional knowledge of the scene. Therefore it would be necessary to integrate standard recognition techniques into the 3D modeling process.

5 Acknowledgments

This work is partly funded by the VRVis Research Ceneter, Graz and Vienna/Austria and the Virtual Heart of Central Europe, Towers, Wells, and Rarities 3D Online project, co-funded by the European Commission in Culture 2000 Framework Programme, Agreement No. 2003 - 1467/ 001 / 001 CLT CA12.

References

BAILLARD, C., SCHMID, C., ZISSERMAN, A., AND FITZGIBBON, A. 1999. Automatic line matching and 3d reconstruction of buildings from multiple views. In *ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery, IAPRS Vol.32, Part 3-2W5*, 69–80.

BAUER, J., KLAUS, A., KARNER, K., SCHINDLER, K., AND ZACH, C. 2002. Metropogis: A feature based city modeling system. *Photogrammetric Computer Vision (PCV)* (September).

BORNIK, A., KARNER, K., BAUER, J., LEBERL, F., AND MAYER, H. 2002. High-quality texture reconstruction from multiple views. In *Journal of Visualization and Computer Animation*.

BROWN, M. Z., BURSCHKA, D., AND HAGER, G. D. 2003. Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*, 8, 993–1008.

DEBEVEC, P. E. 1996. *Modeling and Rendering Architecture from Photographs*. PhD thesis, University

of California at Berkeley, Computer Science Division, Berkeley CA.

FISCHLER, M., AND BOLLES, R. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the Association for Computing Machinery*, 24(6):381–395.

HARRIS, C., AND STEPHENS, M. 1988. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, 147–151.

HORN, B. 1990. Relative orientation. *International Journal of Computer Vision, 4:59–78.*

KLAUS, A., BAUER, J., AND KARNER, K. 2002. Metropogis: A semi-automatic city documentation system. *ISPRS Journal of Photogrammetric Computer Vision* (September), Part A 187–192.

LEYMARIE, F., DE LA FORTELLE, A., KOENDERINK, J., KAPPERS, A., STAVRIDI, M., VAN GINNEKEN, B., MULLER, S., KRAKE, S., FAUGERAS, O., ROBERT, L., GAUCLIN, C., LEVEAU, S., AND ZELLER, C. 1996. Realise: Reconstruction of reality from image sequences. In *IEEE International Conference on Image Processing (ICIP)*, 651–654.

MODELER, R. I., 2004. http://www.realviz.com/products/im/index.php, January.

MORRIS, D., AND KANADE, T. 2000. Image-consistent surface triangulation. In *CVPR 2000*, 332–338.

NISTER, D. 2003. An efficient solution to the five-point relative pose problem. In *CVPR03*, II: 195–202.

POLLEFEYS, M., KOCH, R., VERGAUWEN, M., DEKNUYDT, A. A., AND GOOL, L. J. V. 2000. Three-dimensional scene reconstruction from images. In *Conference on Three-Dimensional Image Capture and Applications II*, SPIE, Bellingham, Washington, B. D. Corner and H. Nurre, Joseph, Eds., 215–226.

REDERT, A., HENDRIKS, E., AND BIEMOND, J. 1999. Correspondence estimation in image pairs. *IEEE Signal Processing Magazine* (May), 29–46.

ROTHWELL, C., MUNDY, J., HOFFMAN, W., AND NGUYEN, V. 1995. Driving vision by topology. In *IEEE Symposium on Computer Vision SCV95*, 395–400.

SCHAFFALITZKY, F., AND ZISSERMAN, A. 2000. Planar grouping for automatic detection of vanishing lines and points. *IVC 18*, 9 (June), 647–658.

SCHMID, C., AND ZISSERMAN, A. 2000. The geometry and matching of lines and curves over multiple views. *IJCV 40*, 3 (December), 199–233.

SCHNEIDERMAN, B. 1998. *Designing the User Interface: Strategies for Effective Human-Computer-Interaction*. Addison-Wesley.

ZARA, J., AND SLAVIK, P. 2003. Cultural heritage presentation in virtual environment: Czech experience. *International Workshop on Database and Expert Systems Applications (DEXA'03).*

ZISSERMAN, A., FITZGIBBON, A., BAILLARD, C., AND CROSS, G. 2000. From images to virtual and augmented reality. In *Conference of Computer Vision and Computer Graphics*, Kluwer Academic Publishers, A. Leonardis, F. Solina, and R. Bajcsy, Eds., NATO Science Series, 1–23. ISBN 0-7923-6612-3.