BUILDING FAÇADE INTERPRETATION FROM IMAGE SEQUENCES

Helmut Mayer and Sergiy Reznik
Institute for Photogrammetry and Cartography, Bundeswehr University Munich, D-85577 Neubiberg, Germany
{Helmut.Mayer|Sergiy.Reznik}@unibw.de

KEY WORDS: Façade Interpretation, Implicit Shape Models, Markov Chain Monte Carlo, Abstraction Hierarchy, 3D Reconstruction

ABSTRACT

In this paper we propose an approach for building façade interpretation ranging from uncalibrated images of an image sequence to the extraction of objects such as windows. The approach comprises several novel features, such as determination of the façade planes by robust least squares matching, learning of implicit shape models for objects, particularly windows, and a determination of the latter by means of a Markov Chain Monte Carlo (MCMC) process employing an abstraction hierarchy. Results for the fully automatic approach show its potential and problems.

1 INTRODUCTION

Automatic interpretation of buildings and particularly their facçades is an area which gained some interest recently. This is illustrated, e.g., by the special issue of "IEEE Computer Graphics and Applications" (Ribarsky and Rushmeier, 2003) comprising, e.g., (Früh and Zakhor, 2003), where a laser-scanner and a camera mounted on a car are employed to generate three-dimensional (3D) models of façades and together with aerial images and laser-scanner data models of areas of cities. Photogrammetrically inspired work focuses on semi-automatic approaches (van den Heuvel, 2001), texturing (Böhm, 2004), and disparity estimation (von Hansen et al., 2004) for façades. Also in the vision community there is interest in the semi-automatic exploitation of the special geometrical constraints of buildings for camera calibration (Wilczkowiak et al., 2005).

Our goal is the automation of the whole process of façade interpretation from video sequences, especially the extraction of objects such as windows. Concerning the detection of façade planes we have been inspired by (Werner and Zisserman, 2002) as well as (Bauer et al., 2003), where Random Sample Consensus – RANSAC (Fischler and Bolles, 1981) as well as plane sweeping is employed. Both detect windows as objects which are situated behind the plane of the façade.

For the extraction of regular configurations of windows, (Wang et al., 2002) present an approach based on oriented region growing taking into account the grid, i.e., row / column, structure of many façades. A more sophisticated approach is given by (Alegre and Dallaert, 2004), where a stochastic context-free grammar is used to represent recursive regular structures on façades. Both models are only demonstrated for one or two rather regular high-rising buildings and it is not really clear, if they are not too strict for general façades.

Of particular interest for our work is (Dick et al., 2004), which is based on a Bayesian model. The basic idea is to construct the building from parts, such as the façades and the windows, changing parameters, e.g., their width, brightness, etc., in a way resulting into an appearance resembling the images. The difference between the model projected into the geometry of the images as well as the prior information on typical characteristics of buildings triggers a statistical process which is implemented in the form of Reversible Jump Markov Chain Monte Carlo - RJMCMC (Green, 1995). RJMCMC is used as it can deal with the fact, that the number of objects changes during processing. We also integrate prior as well as image information, but as we do not change the number of object instances (yet), we use traditional Markov Chain Monte Carlo – MCMC (Neal, 1993).

In Section 2 we sketch our approach to generate a Euclidean 3D model from uncalibrated image sequences before determining the vertical vanishing point, the façade planes, as well as points lying on them (cf. Section 3). Section 4 shows how image patches around interest points can be used to learn an implicit shape model for windows which is employed to extract hypotheses for windows. The latter are used to extract windows by means of MCMC on an abstracted version of the original image generated by means of a Dual Rank filter (cf. Section 5). The paper ends up with conclusions.

2 3D RECONSTRUCTION AND CALIBRATION

Our approach for 3D reconstruction and calibration is aiming at full-automation for wide-baseline image sequences of rather large images. Therefore, we employ image pyramids and sort out blunders via RANSAC and fundamental matrix as well as trifocal tensor (Hartley and Zisserman, 2003). The latter is based on highly precise conjugate points derived from Förstner points (Förstner and Gülch, 1987). If the (normalized) cross-correlation coefficient (CCC) is above a relatively low threshold, we determine the sub-pixel precise shift by means of affine least squares matching of all corresponding image patches.

We start by generating image pyramids, with the highest pyramid level in the range of about 100×100 pixels. On this level we determine point pairs and from them fundamental matrices $\mathbf{F}$ for all consecutive pairs. The epipolar lines derived from $\mathbf{F}$ guide the matching of triplets on the

second highest pyramid level which lead to trifocal tensors $\mathcal{T}$. With $\mathcal{T}$ we filter out most blunders. After determining $\mathbf{F}$ as well as $\mathcal{T}$ with the usual linear algorithms (Hartley and Zisserman, 2003) we do a robust, at this stage projective bundle adjustment. If the image is larger than about 1000×1000 pixels, $\mathcal{T}$ is also determined on the third highest pyramid level.

Each triplet linked by $\mathcal{T}$ has its own 3D projective coordinate system. To link the triplets, we use the direct linear transformation (DLT) for points mapped from the preceding into the current triplet based on $\mathcal{T}$. Additionally, we integrate points into the solution, which could not be seen in preceding triplets. After linking the triplets and including new points we again compute a robust projective bundle adjustment. When all triplets have been linked on the second or third highest pyramid level, we track all points through the pyramid by robust least squares matching in all images. This results into sub-pixel coordinates for all points in relation to a master image on the original image size. The points are input to a final (projective) bundle adjustment including radial distortion.

To obtain the internal camera parameters, we use the approach proposed in (Pollefeys et al., 2004) based on the image of the absolute dual quadric Ω^*. For the latter holds $\omega^* \sim \mathbf{P}\Omega^*\mathbf{P}^\top$, with $\mathbf{P}$ the projection matrix for the i-th camera and $\omega \sim \mathbf{K}\mathbf{K}^\top$, $\mathbf{K}$ being the calibration matrix comprising the internal parameters principal distance and point as well as scale difference and skew. The idea of (Pollefeys et al., 2004) is to impose constraints on the internal parameters, such as, that the (normalized) principal distance usually is one with a standard deviation of, e.g., three, the principal point is close to the center, the skew is very small, and there is only a small scale difference. From it we obtain in most cases a meaningful solution which is then finally polished via robust Euclidean bundle adjustment.

Results for orientation and reconstruction consisting of about 450 3–fold and 370 4–fold points can be seen on the right hand side of Figure 1, showing on the left three images from Prague's famous Hradschin. The right angle at the building corner has been reconstructed rather well.

3 DETERMINATION OF FAÇADE PLANES AND THE POINTS ON THEM

Before generating façade planes, we take into account one of the most general constraints for façades, namely being oriented vertically. With it, we can later safely assume, that windows or doors are rectangles oriented in parallel to the coordinate axes. The basic idea is, that all vertical lines are parallel in space, their projections in an image therefore intersecting in a specific vanishing point. Usually, the vertical vanishing point is, depending on holding the camera upright or rotated 90°, in the y- or in the x-direction. As it is difficult to decide from the image alone, in which of these two directions the vertical vanishing point actually lies, we input this information by means of a flag, telling that the vanishing point is more in y- (standard) or x-direction. Everything else is done automatically.

We start by extracting straight lines with the Burns-operator. Hypotheses for vanishing points are found by means of RANSAC and supporting lines are used to improve the coordinates of the vanishing point via least squares adjustment. From the best hypotheses we take the one, which is closest to the direction in which we know the vertical vanishing point should be. An example is given in Figure 2. Knowing the vanishing point and the calibration matrix $\mathbf{K}$ from the preceding section, we can directly compute the vertical direction in space. To improve the quality, we compute the vertical vanishing point for more than one image, relate the results via the known orientation parameters of the cameras, and then compute the average.

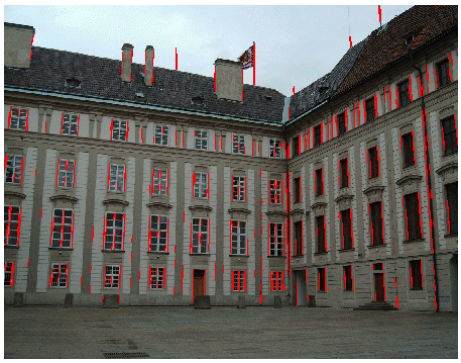

Figure 2: Lines defining the vertical vanishing points

Hypotheses for façades are generated in the form of planes from the (Euclidean) 3D points generated as a result of the preceding section. To filter out the points on the planes, we again employ RANSAC. A plane can be parameterized by three parameters in the form of a homogeneous 4-vector and can be determined accordingly from three points. We randomly take three points, determine a plane from them, and then check how many points are close to that plane. The latter needs one threshold, which depends on factors such as the resolution of the camera, the actual planarity of the plane (old buildings might be less planar, if at all), and the geometry of the acquisition configuration. Therefore, it is justified, to optimize it by hand.

Opposed to standard RANSAC, we do not just take the best solution in terms of the number of points on the plane, but the set of all mutually only little overlapping hypotheses, starting with the best hypothesis. This is because there might be more than one planar façade in the scene and the corresponding planes might have common points on intersection lines. The latter motivates an allowed overlap of several percent.

The obtained (infinite) planes are restricted by means of the bounding rectangle of all points on the plane, taking into account the known vertical direction. To further restrict the points (pixels) on the façade and to improve the parameters of the plane, we use robust least squares matching. Knowing the projection matrices for the cameras as well as the plane parameters, we compute homographies. They

Figure 1: Three images of wide-baseline quadruple "Prague" and result after orientation and calibration (points in red, cameras as green pyramids; $\sigma_0 = 0.24$ pixels)

allow us to transform the information supposed to be on the given plane from all images into the same image geometry. Figure 3 shows on the left three of the four images of the Prague scene projected in black-and-white onto one of the façade planes, mapped into the red, green, and blue channels. If all pixels were on the plane and there was no radiometric difference between the images, the combined image should be black-and-white. Colors therefore show deviations from the plane but also in radiometry. This fact is employed by a least squares optimization of the three plane parameters including the elimination of outlier pixels, implicitly classifying the points on the plane. The result for this is given on the right of Figure 3, the black parts lying on the façade plane, while white holes can be seen as hypotheses for windows, doors, or other architectural elements. Figure 4 shows both dominant planes computed from about 270 and 250 supporting points, respectively, including the holes and the 3D points.

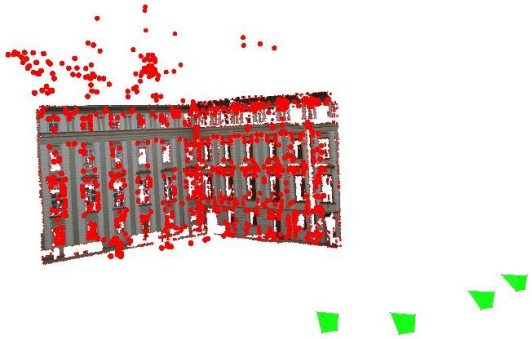

Figure 4: The two dominant planes restricted to the areas classified to be on the plane together with the 3D points (red) and the cameras (green pyramids)

4 DETECTION OF WINDOWS BASED ON AN IMPLICIT SHAPE MODEL

As can be seen from Figure 3, the detection of holes in the regions corresponding to a façade plane comprises one possible means to hypothesize windows. Yet, it is not extremely reliable, as windows tend to be dark with low contrast not generating outliers, i.e., holes. Therefore, we have devised another means to generate hypotheses for windows based on ideas put forward by (Agarwal et al., 2004) and (Leibe and Schiele, 2004). Basically, use is made of the shape of an object in the form of the arrangement or the relations of characteristic parts, e.g., image patches. As (Agarwal et al., 2004) and (Leibe and Schiele, 2004) we use CCC to decide, if image patches are similar. While (Agarwal et al., 2004) learn the angle and distance between image patches clustered together based on the CCC to find cars in ground-based images, (Leibe and Schiele, 2004) employ a generalized Hough transform.

We follow the latter idea and assume that the images have been projected onto the façade plane, are oriented by knowing the direction of the vertical vanishing point, and have been scaled by one factor to approximately the same scale ($\pm$ about 20%). Instead of clustering the image patches, we simply "learn" the shape of a window as follows (this can be seen as a simplified version of (Leibe and Schiele, 2004)): We cut out image parts around windows. In these we extract Förstner points with a fixed set of parameters, mark by hand the center of the window, and then we store the difference vectors between the points and the center as well as image patches of size 13×13 pixels around the points. Eleven out of 72 windows used for training resulting into 702 points are given in Figure 5.

To detect windows on a façade, we extract Förstner points with the same set of parameters as above and compare the patches of size 13×13 centered at them with all points learned above by means of CCC. If the latter is above a threshold of 0.8 found empirically, we write out the difference vector for the corresponding point into an initially empty evidence image, incrementing the corresponding pixel by one. I.e., each point (possibly multiply) votes for the position of the window center. The points on the façade as well as the image array with the evidence for the position of the window centers are given for our running example in Figure 6.

Figure 6 shows, that the evidence for the window centers is widely spread, because some parts of the windows vote for different positions. This is due to the fact, that a patch can look, e.g., similar to an upper right corner of a whole window but also of a window part. To obtain meaningful hypotheses, we integrate the evidence for the centers by smoothing them with a Gaussian and then determine all maxima above a threshold. The result for this is given, e.g., in Figure 7, showing reasonable hypotheses. Please note, that none of the windows used for training stems from this scene.

Figure 3: Three black-and-white images projected onto one of the façade planes mapped into the red, green, and blue image channel showing radiometric differences, but also deviations from the plane (left) and pixels on the plane (right)

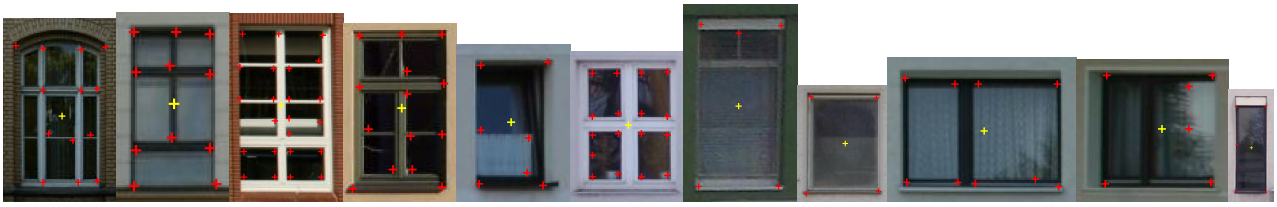

Figure 5: Eleven out of 72 windows used for training with Förstner points (red) and window center (yellow)

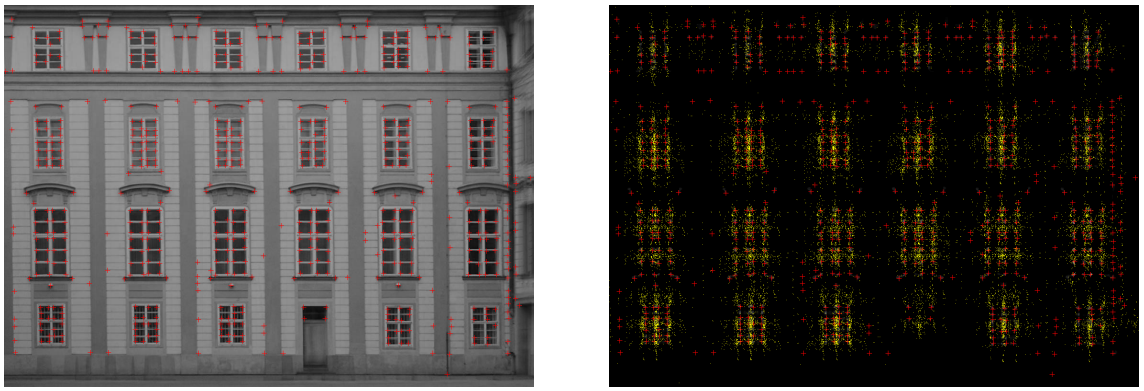

Figure 6: Façade (left) and accumulated evidence for centers of windows (right), both with Förstner points (red)

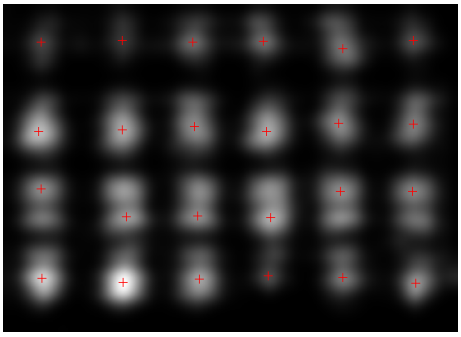

Figure 7: Accumulated evidence for window centers integrated with Gaussian and maxima (red cross)

5 MCMC BASED EXTRACTION OF WINDOWS

Our extraction scheme for the extent of windows is based on the following assumptions / experiences:

- Window panes mostly appear dark during the day when images are taken. This is particularly true for the red channel, because windows consist of glass, which is more easily passed by red light, and because the sky reflected in the windows is mostly blue.
- Most windows are at least partially rectangular with one side being vertical to be able to open the window easily.
- Studying a large number of windows showed that the ratio of height to width of a window lies in the majority of cases between 0.25 to 5. Very narrow windows are encountered more frequently than very wide.
- Windows are often complex objects consisting of different parts such as window-sills, mullions, transoms, and sometimes also flower pots.

The latter can also be interpreted in terms of abstraction by means of a scale-space. What we are interested into is an object in the range of about 1×1.5 meter width and

height, but not the smaller details. To get rid of them, we generate an abstract version of the object by means of a suitable scale-space. One scale-space which was proven to give meaningful results for this kind of problem, where objects have a stark contrast, is gray-scale morphology in the form of opening and closing. It can be made more robust by not taking the infimum or supremum, but, e.g., the 5% quantile, and is then termed Dual Rank filter in (Eckstein and Munkelt, 1995). Here, we use opening with a radius of about 10 cm eliminating dark parts followed by closing with a radius of about 25 cm eliminating bright parts (cf. Figure 8). The opening before the closing is necessary to avoid, that bright parts cannot be get rid of because they are disturbed by small dark parts.

To actually extract windows, we take up the basic idea of (Dick et al., 2004) and try to generate an image which is at least in some respect similar to the actual image. Our model is very simple, namely dark rectangles on a bright background. This can be seen as the third level of an abstraction hierarchy consisting additionally of the original image, and the Dual Rank filtered image (cf. Figure 8).

The model is disturbed by Gaussian noise and compared to the actual façade image by means of CCC. For each iteration of MCMC, we either change the width, the height, or the position of the dark rectangle representing the window. The probability is 30% for a change of width or height and 20% for a change of the horizontal or vertical position, respectively. It reflects our assumption, that we know more about the position than about the size. This is natural for a hypotheses stemming from a procedure determining only the center of a window, though we know the average sizes of windows. To robustify the search, we use simulated annealing. I.e., the higher the number of iteration becomes, the lower becomes the probability to accept results which are worse than for the preceding iteration. To optimize the process, we do not compare the whole façade with a window but only a rectangular image part five times larger than the average window size.

Figure 9 shows the result of the above process. Hypotheses were generated in the form of relatively small squares at the positions of the maxima of the implicit shape model based approach proposed above, leading to a fairly reasonable result. Further results are given in Figure 10. For both there is room for improvement for the delineation of the windows.

Figure 9: Result of MCMC – hypotheses (white box) and windows (green box)

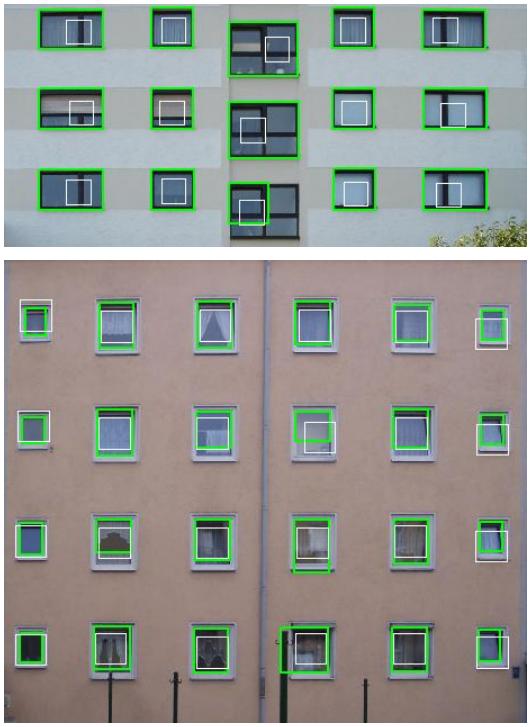

Figure 10: Further results – hypotheses (white box) and windows (green box)

6 CONCLUSIONS

The results we have presented have been produced fully automatically, using only very few semantically meaningful thresholds, such as the planarity of walls. Yet, there is ample room for improvement. One possible way to pursue would be to make more use of the geometric regularity of the scene, e.g., for camera calibration. We will focus on the integration of the learned implicit shape model into the MCMC process, using points at different image scales, i.e., abstraction levels. We assume, that for this we will need to form clusters for the image patches in the same way as (Agarwal et al., 2004, Leibe and Schiele, 2004). We also want to make use of the fact, that façades often consist of regular structures in the form of rows and columns, which will model more explicitly the abstraction hierarchy of façades. For this, it will be necessary to be able to in- and exclude objects in the statistical process by means of RJMCMC. Finally, on a wider time scale, we want to model façades in 3D, by matching the obtained window hypotheses in the images, e.g., by sweeping, but also by including prominent 3D objects such as balconies.

ACKNOWLEDGMENTS

Sergiy Reznik is funded by Deutsche Forschungsgemeinschaft under grant MA 1651/10. We thank the anonymous

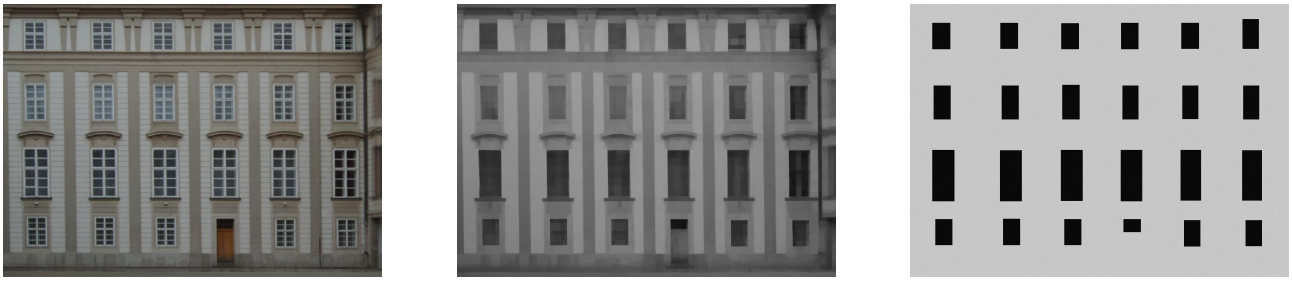

Figure 8: Abstraction hierarchy consisting of the original image (left), the Dual Rank filtered image (center), and the MCMC model (right)

reviewers for their helpful comments.

REFERENCES

Agarwal, S., Awan, A. and Roth, D., 2004. Learning to Detect Objects in Images via a Sparse, Part-Based Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(11), pp. 1475–1490.

Alegre, F. and Dallaert, F., 2004. A Probalistic Approach to the Semantic Interpretation of Building Facades. In: International Workshop on Vision Techniques Applied to the Rehabilitation of City Centres, pp. 1–12.

Bauer, J., Karner, K., Schindler, K., Klaus, A. and Zach, C., 2003. Segmentation of Building Models from Dense 3D Point-Clouds. In: 27th Workshop of the Austrian Association for Pattern Recognition.

Böhm, J., 2004. Multi Image Fusion for Occlusion-Free Façade Texturing. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. (35) B5, pp. 867–872.

Dick, A., Torr, P. and Cipolla, R., 2004. Modelling and Interpretation of Architecture from Several Images. International Journal of Computer Vision 60(2), pp. 111–134.

Eckstein, W. and Munkelt, O., 1995. Extracting Objects from Digital Terrain Models. In: Remote Sensing and Reconstruction for Three-Dimensional Objects and Scenes, 2572, SPIE, pp. 43–51.

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 ACM 24(6), pp. 381–395.

Förstner, W. and Gülch, E., 1987. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In: ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data, Interlaken, Switzerland, pp. 281–305.

Früh, C. and Zakhor, A., 2003. Constructing 3D City Models by Merging Aerial and Ground Views. IEEE Computer Graphics and Applications 23(6), pp. 52–61.

Green, P., 1995. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika 82, pp. 711–732.

Hartley, R. and Zisserman, A., 2003. Multiple View Geometry in Computer Vision – Second Edition. Cambridge University Press, Cambridge, UK.

Leibe, B. and Schiele, B., 2004. Scale-Invariant Object Categorization Using a Scale-Adaptive Mean-Shift Search. In: Pattern Recognition – DAGM 2004, Springer-Verlag, Berlin, Germany, pp. 145–153.

Neal, R., 1993. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K. and Tops, J., 2004. Visual Modeling with a Hand-Held Camera. International Journal of Computer Vision 59(3), pp. 207–232.

Ribarsky, A. and Rushmeier, H., 2003. 3D Reconstruction and Visualization. IEEE Computer Graphics and Applications 23(6), pp. 20–21.

van den Heuvel, F. A., 2001. Object Reconstruction from a Single Architectural Image Taken with an Uncalibrated Camera. Photogrammetrie – Fernerkundung – Geoinformation 4/01, pp. 247–260.

von Hansen, W., Thönnessen, U. and Stilla, U., 2004. Detailed Relief Modeling of Building Facades From Video Sequences. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. (35) B3, pp. 967–972.

Wang, X., Totaro, S., Taillandier, F., Hanson, A. and Teller, S., 2002. Recovering Facace Texture and Microstructure from Real-World Images. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. (34) 3A, pp. 381–386.

Werner, T. and Zisserman, A., 2002. New Techniques for Automated Architectural Reconstruction from Photographs. In: Seventh European Conference on Computer Vision, Vol. II, pp. 541–555.

Wilczkowiak, M., Sturm, P. and Boyer, E., 2005. Using Geometric Constraints through Parallelepipeds for Calibration and 3D Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(2), pp. 194–207.