Photogrammetrie • Fernerkundung • Geoinformation 3/2007, S. 167–176, 8 figs.

3D Reconstruction and Visualization of Urban Scenes from Uncalibrated Wide-Baseline Image Sequences

HELMUT MAYER, Neubiberg

Keywords: Photogrammetry, 3D reconstruction, auto-calibration, markerless orientation, visualization

Summary: This paper focuses on the fully automatic generation of basic ingredients for high quality visualizations of urban areas characterized by vertical facade planes. We show that uncalibrated wide-baseline image sequences without using markers or ground control suffice for this task. At the core of our algorithms are least-squares matching, projective geometry based reconstruction, robust estimation based on random sample consensus – RANSAC, direct auto-calibration, projective and Euclidean bundle adjustment, plane to plane homographies, as well as the robust estimation of image mosaics. Results for the Hradschin in Prague, Czechia, Plaza Real in Barcelona, Spain, and the Zwinger in Dresden show the potential and shortcomings of the employed algorithms.

Zusammenfassung: *3D Rekonstruktion und Visualisierung von städtischen Szenen auf der Grundlage von unkalibrierten Bildsequenzen mit großer Basis*. Dieses Papier zielt auf die vollautomatische Generierung von grundlegenden Bestandteilen für hochqualitative Visualisierungen von städtischen, durch vertikale Fassadenebenen charakterisierte Szenen ab. Es wird gezeigt, dass für diese Aufgabe unkalibrierte Bildsequenzen mit großer Basis ohne Verwendung von Messmarken oder Passpunkten ausreichen. Den Kern der vorgestellten Algorithmen bilden kleinste-Quadrate-Zuordnung, Rekonstruktion auf Grundlage projektiver Geometrie, robuste Schätzung basierend auf random sample consensus – RANSAC, direkte auto-Kalibrierung, projektive und euklidische Bündelausgleichung, Ebene-zu-Ebene Homographien, sowie die robuste Schätzung von Bildmosaiken. Ergebnisse für den Hradschin in Prag, Tschechien, den Plaza Real in Barcelona, Spanien und den Zwinger in Dresden zeigen die Möglichkeiten aber auch die Defizite der verwendeten Algorithmen.

1 Introduction

Microsoft recently announced its Photosynth project (http://labs.live.com/photosynth/). Right now users can only view colored Euclidean three-dimensional (3D) point sets and images registered to average planes of 3D scenes. Yet, the project aims at that the user can include her/his photos of an uncalibrated camera, orient them in relation to the given 3D point sets and possibly also extend the 3D point sets. We are following a similar trail, but we restrict ourselves to high precision 3D points and vertical planes in an urban setting.

Recent years have seen a couple of approaches for the fully automatic generation of 3D Euclidean models from uncalibrated image sequences, among the most advanced of which is (POLLEFEYS et al. 2004). The approaches usually consist of the robust estimation of a projective reconstruction and n-fold correspondences followed by auto-calibration and possibly dense depth estimation, all usually restricted to small images with a short baseline, e. g., from a video camera.

1432-8364/07/2007/0167 $ 2.50

Opposed to this, we aim at applications where higher image resolutions in the range of several Megapixels are given as input, obtained, e. g., from consumer digital cameras costing only several hundred Euros. Because of the lower frame rates (one image can usually be taken on a sustained basis only about every second on average) and higher data volumes per image it is natural to take images with a wider baseline making the matching of points between the images severely more difficult. We show how employing high precision to become more reliable it is possible to obtain 3D reconstructions of rather difficult scenes with many occlusions and partly close to no 3D structure.

The focus of this paper is on urban scenes. Therefore, it is reasonable to use at least partly planes for the modeling and visualization of the scenes, particularly the vertical planes of the facades. As larger parts of our scenes are assumed to be captured in at least three images, it becomes on one hand necessary to fuse the information from the individual images on the detected planes. Yet, on the other hand, it gives us the opportunity, to separate by means of consensus between pixels taken from different images the information on the plane from off-plane information. This allows us to generate a ''cleaned'' version of the image on the plane without many of the occlusions in the individual images. This part has been inspired by (Böhm 2004). Yet, opposed to the latter, we fully automatically and robustly generate the planes and from them the two-dimensional (2D) homographies, i. e., plane to plane mappings. We also integrate the planes into our 3D models and generate visualizations from them.

Impressive results in terms of visualization of urban scenes have been presented by Debevec et al. (1996) by taking the image from the (real) camera closest to the current (virtual) viewpoint. Yet, the 3D model employed has been generated manually. Then, there is work for architectural scenes which goes far beyond what we are presenting here in the sense that much more knowledge about the structures and regularities of urban scenes is used. The most sophisticated example is probably (Dick et al. 2004) employing a statistical generative model based on Markov Chain Monte Carlo (MCMC) sampling. Closer to our work as it is more geometry-based is (Werner & Zisserman 2002). Yet, compared to our work they employ perpendicular vanishing points for auto-calibration and 3D reasoning which restricts the work to scenes with three perpendicular main directions. They have also only shown results for image triplets.

In the remainder of this paper, we first present our approach for 3D reconstruction from wide-baseline image sequences (cf. Section 2). The obtained 3D Euclidean model is the basis for deriving vertical facade planes. For them facade images at least partly ''cleaned'' from occlusions are computed by means of median or consensus between the pixels projected onto the planes from different camera positions (cf. Section 3). In Section 4 we present additional results and we end up with conclusions.

2 3D Reconstruction

Our approach aims at wide-baseline image sequences made up of images of several Megapixels. We make the following assumptions for 3D reconstruction:

- The camera constant (principal distance) is constant. Yet this is not as restrictive as it may sound because we found that the influence of auto-focusing that one cannot switch off for some cameras we use can mostly be neglected for the distances typical for urban applications. We also assume that the principal point is close to the image center. This is the case for practically all digital cameras, and would only not hold if parts of images were used.
- The images are expected in the form of a sequence with at least three-fold overlap for all images.

Our basic idea to obtain a reliable result is to strive for a very high precision in the range of 0.05 to 0.3 pixels by means of least-squares matching and bundle adjustment. If the value is higher or lower depends in first

instance on scene geometry and geometrical quality/stability of the camera, but in second instance also on lighting conditions, etc. The overall reasoning is that it is (extremely) unlikely that a larger number of non-homologous points conspire to achieve a highly precise result by chance.

Based on this idea we start using Förstner points (Förstner & Gülch 1987). They are matched via cross-correlation. In color images the coefficient for the channel where the variance is maximum is taken. To deal with images rotated around the axis of the camera, we rotate each patch according to the direction obtained for the Förstner points. From the points accepted by matching we compute a histogram of the relative directions between the matched points. The angle for which the histogram is maximum is used to rotate all patches of the second image according to the reference image. Point pairs checked via correlation are refined via least-squares matching with an affine geometrical model. The latter is also used for three- and more-fold images. In all cases we compute the complete covariance information.

The highly precise points are the basis for a projective reconstruction employing fundamental matrices F and trifocal tensors T (Hartley & Zisserman 2003). If calibration information is available, we use (Nistér 2004) to determine the Euclidean 3D structure for image pairs. As in spite of our efforts to obtain reliable matches we obtain partly less than 10 % of correct homologous points for difficult scenes, we employ Random Sample Consensus – RANSAC (Fischler & Bolles 1981) for the estimation of F and T. Because we do not only have rather low numbers of correct matches (inliers), but as these inliers are also partly very unevenly distributed over the image and thus not all of them lead to a correct model, i. e., a model representing all inliers with the inherent, yet unknown achievable geometric accuracy, we employ a variant of the locally optimized RANSAC scheme of Chum et al. (2003). While they take a larger number, i. e., 50 %, of random samples from the maximum set of inliers derived at a certain stage to derive an improved estimate, we take the whole maximum set and employ robust bundle adjustment (Hartley & Zisserman 2003, Mikhail et al. 2001). The latter is done two times always using the outcome of the bundle adjustment to derive new sets of inliers.

The employed bundle adjustment is suitable for the projective as well as the Euclidean case. We model radial distortion with a quadratic and a quartic term. Bundle adjustment takes into account the full covariance information derived by least-squares matching. We estimate the precision of the residuals and use them in two ways to make the adjustment robust: First, we reweight the observations based on the ratio of the size of the residual and its variance. Second, after convergence we throw out all points with a ratio beyond three, a value found empirically.

As our images are in the range of several up to possibly tens of Megapixels, it is important to initially constrain the search space for matching. Yet, because we do not want to constrain the user more than given in the assumptions at the begin of the section, we cannot assume that the movement is only vertically or horizontally or that it is even in a certain range. Particularly for urban scenes with very close and far away objects disparities can be rather large, in the extreme case the image size. We thus take as initial search space the full image, but reduce the image in a pyramid and do the first search on a pyramid level with a size of approximately 100×100 pixels. Here, full search can be done efficiently. Matching and projective reconstruction lead to fundamental matrices and thus epipolar lines on the highest level, restricting the search on the next level considerably. Once trifocal tensors have been determined, the search space becomes a small area in the third image. Trifocal tensors are computed for the second highest level in all cases and additionally on the third highest level if the image size exceeds one Megapixel.

To orient whole sequences, we link triplets based on 3D homographies computed from projection matrices for images common be-

Fig. 1: Six images of the Hradschin in Prague, Czechia.

Fig. 2: 3D points, colored according to the pixels, and cameras (green pyramids, the tip symbolizing the projection center and the base giving the direction of the camera) derived from the images given in Fig. 1.

tween triplets. (E. g., the triplets (1,2,3) and (2,3,4) have the images 2 and 3 in common.) Additionally, we project already known 3D points into the newly linked image to generate $i+1$-fold points, with i being the current number of images a point is visible in. After these steps we bundle adjust the sequence. Once all projection matrices and 3D points have been computed, we track the points generated on the second or third highest level of the pyramid down to the original resolution again via least-squares matching in all images.

If no calibration information is given, we directly auto-calibrate the camera employing the approach proposed by Pollefeys et al. (2004), It uses only very weak and general information about cameras to constrain the solution, e. g., that the principal point corresponds to the center of the image and that the camera constant is somewhere in-between one-third and three. Auto-calibration is done only once a high quality projective reconstruction has been obtained on the original resolution via projective bundle adjustment. We found that the latter is mandatory, as lower precisions lead to incoherent implicit calibrations of the projective reconstructions, often leading to unacceptable results. Finally, we employ Euclidean bundle adjustment to obtain a highly-precise calibrated 3D model consisting of points and projection matrices including full covariance information.

If the image sequence consists of a loop, i. e., the first and the last images are the same, we extend the end of the sequence up to the second image. By this means, we get a 3D overlap between the begin and the end which we use to close the loop, thus avoiding a gap between last and first image and evenly distributing the deformation of the whole sequence by error propagation.

An example is given in Fig. 1 and 2 showing a part of the Hradschin in Prague, Czechia. The back-projection error of the calibrated bundle is $\sigma_0 = 0.16$ pixels in the given 2 Megapixel images. Several hundred six-fold points have been computed. One can see that the right angles in the center of

the building have been derived very accurately.

3 Planes and Images on Planes

We assume that an urban scene consists of a considerable number of vertical lines. We can thus orient the 3D Euclidean model vertically based on the vertical vanishing point derived from the vertical lines and the given calibration information. The vertical vanishing point is robustly detected again using RANSAC, the user only providing the information if the camera has been been very approximately held horizontally or vertically, thus, avoiding to mix up the vertical with a horizontal vanishing point. After detecting the vanishing point, we polish it by means of least-squares adjustment. To make the computation of the vertical direction more robust, we compute vanishing points for a couple, usually if possible five images, derive from all of them the vertical direction of the whole model employing the known rotation of the individual camera, and then finally take the medians in x- and y-direction as the vertical direction.

The vertically oriented model is the basis for the determination of vertical facade planes once again using RANSAC. For this step one threshold defining the maximum allowed distance of points from the plane has to be given by the user. This is due to the fact that we could determine meaningful thresholds for approximating planes from the covariance matrices via model selection, but this would only take into account the measurement accuracy and not the semantically important construction precision of facade planes.

To make it more robust and precise, we employ the covariance information of the 3D points computed by bundle adjustment by not counting the number of inliers as for standard RANSAC, but testing the distances to a hypothesized plane based on the geometric robust information criterion – GRIC (Torr 1997). Additionally, we check if the planes are at least approximately vertical and we allow only a limited overlap of about five percent between planes. The latter is needed, because of points situated on intersection lines between planes.

From the parameters for the facade planes as well as the projection matrices we compute homographies between the planes and the images. A mapping by a homography H between homologous points x and x' in homogeneous coordinates on a given plane and the image plane of a camera, respectively, is given by

$$\mathrm{x}' = \mathrm{H}\mathrm{x}. \tag{1}$$

The camera is parameterized as

$$P = \begin{pmatrix} P_{11}P_{12}P_{13}P_{14} \\ P_{21}P_{22}P_{23}P_{24} \\ P_{31}P_{32}P_{33}P_{34} \end{pmatrix} \tag{2}$$

and the plane, the points lie on, with the four-vector

$$\pi = (\mathbf{n}^{\mathrm{T}}, d)^{\mathrm{T}}. \tag{3}$$

We parameterize the plane in 2D by setting that component of the first three components of the plane with maximum value to zero. By this means we obtain the mapping in the direction closest to the normal. We index the maximum value with m and the other two with j and k. Then H is determined as ($i \in 1,2,3$)

$$\mathsf{H} = \begin{pmatrix} H_{i1} \\ H_{i2} \\ H_{i3} \end{pmatrix} = \begin{pmatrix} P_{ij} - \pi_j \cdot P_{im}/\pi_m \\ P_{ik} - \pi_k \cdot P_{im}/\pi_m \\ P_{i4} - \pi_4 \cdot P_{im}/\pi_m \end{pmatrix}. \tag{4}$$

For the actual mapping of images to a plane one needs to know from which images a plane can be seen. For it, the information is employed, which 3D points have led to a particular plane, as for the 3D points it is known from which images they were derived. The plane is thought to be visible from the union of the sets of images of all 3D points belonging to a plane. We compute an average image as well as the bias in brightness for each image in comparison to it, also accounting for radial distortion.

The final step is the generation of facade images if possible "cleaned" from artifacts generated by occlusions. The basic informa-

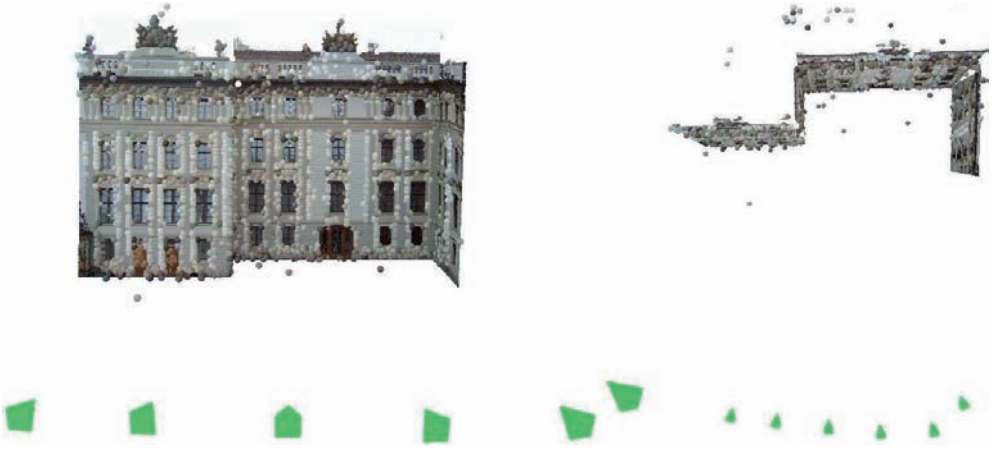

Fig. 3: 3D points, colored according to the pixels, facade planes and cameras (green pyramids; cf. Fig. 2) derived from the images given in Fig. 1 and the 3D model in Fig. 2.

Fig. 4: Facade image derived from the six images given in Fig. 1 – left: average; center: median; right: consensus.

tion are the projected images normalized via the determined biases in brightness. The cleaning is done by two means, first by sorting the (gray- or color) values and taking the median and second by utilizing the basic idea of Böhm (2004). The latter consists in determining an optimum value by means of the consensus between the values for a particular pixel. As Böhm (2004) we do not randomly select the values as in RANSAC, but we take the value for a pixel for each image it can be seen from as estimate and then take as the inliers all values which consent with it. The final result is the average of the inliers.

Results for our running example are given in Fig. 3 and 4. From the former one can see that the planes nicely fit to the points. The latter shows the advantages of median and consensus over simple averaging where, e. g., the flag pole at the right hand side is shown several times as a ghost image. The different characteristics of median and consensus are shown more in detail in the additional example in the next section.

4 Additional Results

We took a set of 29 uncalibrated images of Plaza Real in Barcelona, Spain with a Sony P 100 5 Megapixel camera. The basic idea was to walk around the fountain in the center of Plaza Real. A 3D model is computed (cf. Fig. 5) with $\sigma_0 = 0.18$ pixels after bundle adjustment. As we did not mark our positions when taking the images, the circle around the fountain is more a spiral. Opposed to previous results published in MAYER, we now could close the loop. The right angles have been determined very well in spite of the relatively large areas where we could not match due to occlusions mostly by the palm trees.

The facade image for the left facade in Fig. 5 derived from the ten images shown in Fig. 6 is given in Fig. 7. First, the average image shown at the bottom makes clear by means of the circular streaks how large the influence of radial distortion is for some of the images. (Please note that the images with the largest distortions look onto the plane from the side, strongly amplifying the effect.) Overall, one can see that the average is not acceptable. This is due to the ghost images of the occluding objects, but also because of a not precise enough estimation of the bias of the brightness between the average image and the individual images. The latter stems from the unmodeled occlusions which lead to estimating wrong biases from pixels representing different objects. The latter problem could only be dealt with by robustly recursively estimating biases and occluding objects, which is non-trivial and on our agenda for further research.

Opposed to the average, the median and the consensus do much better, even though both are not able to penetrate the vegetation in many instances. If the vegetation is dense, this is not possible at all, but the problem could partly be alleviated by means of more images from different positions. Concerning median and consensus, there are only some, yet characteristic differences. One of the largest can be seen left of the center. The first leaf of the palm tree is mostly eliminated by

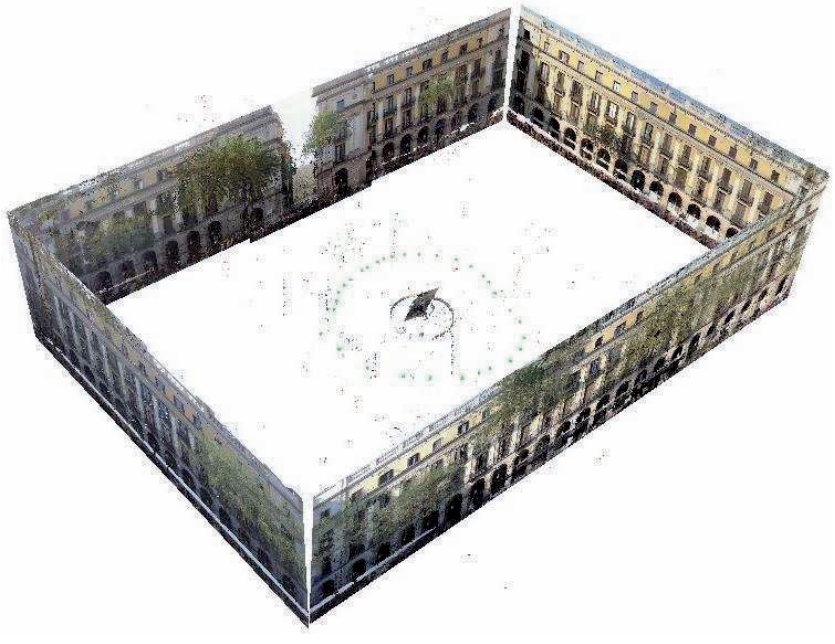

Fig. 5: 3D points, facade planes, and cameras (medium sized circle around the fountain in the center) derived from nothing but uncalibrated images, ten of them showing the facade on the left hand side given in Fig. 6.

the consensus, but not by the median, as the former has a weaker basic assumption and can thus deal with more than 50 % of outliers.

In Fig. 8 an additional example is presented generated from ninety images of the Zwinger in Dresden. It shows shortcomings of using only planes for modeling the surfaces. The left and the right part consist of curved surfaces and the gates on the left and the right side are highly 3D structured and thus cannot be adequately modeled by planes.

Fig. 6: Ten images of Plaza Real in Barcelona from which the facade images given in Fig. 7 have been derived.

Fig. 7: Facade image derived from the ten images given in Fig. 6 – top: consensus; center: median; bottom: average.

5 Conclusions

We have shown how combining projective reconstruction with robust techniques and bundle adjustment propagating covariance information can be used to fully automatically generate textured 3D models of urban scenes from nothing but (possibly uncalibrated) perspective images also for larger numbers of wide-baseline images. These still incomplete 3D models can be the basis for high quality visualizations. Though, at the moment lots of additional manual efforts are needed for a practically satisfying outcome.

One way to proceed is to add detailed geometry by employing semantic information, e. g., by 3D extraction of the win-

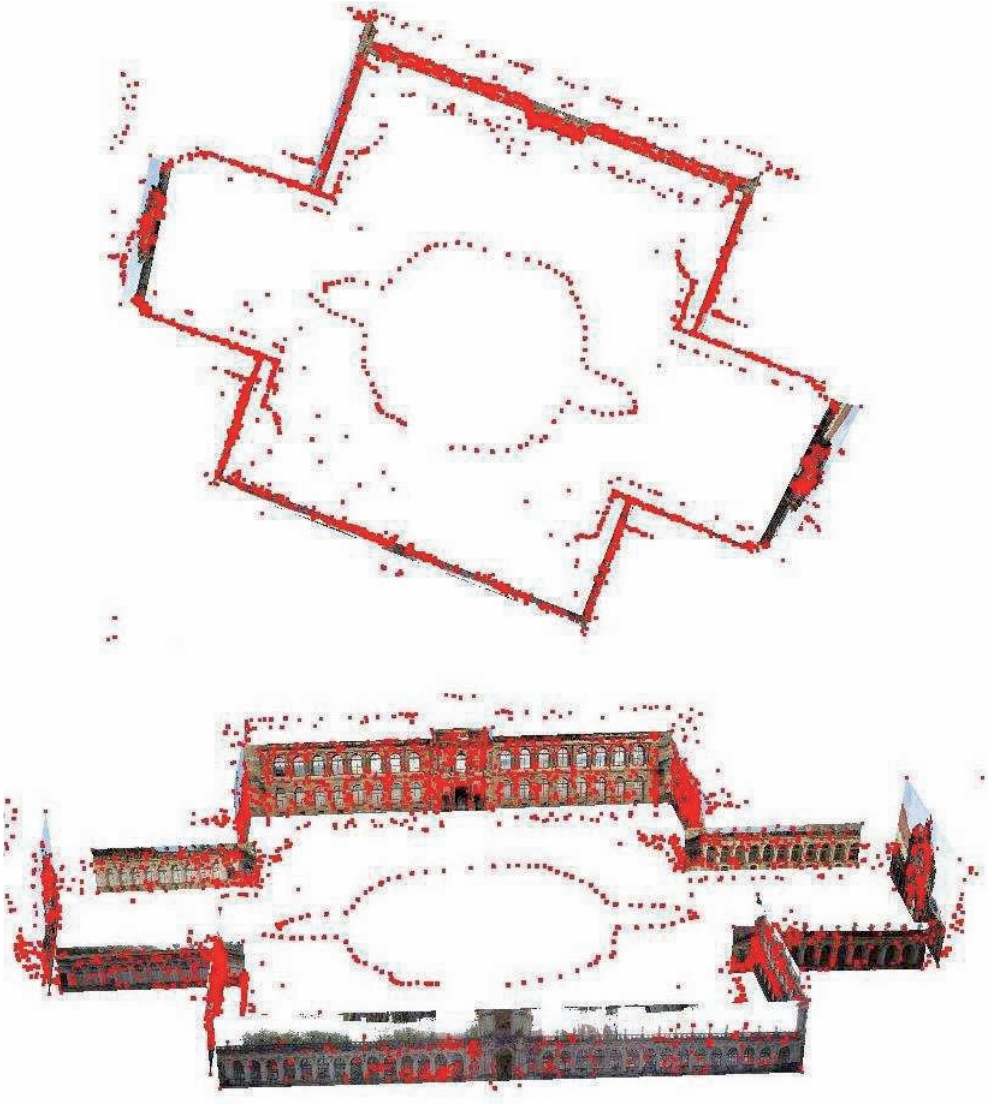

Fig. 8: 3D model of Zwinger, Dresden: 3D points and cameras in red – the 90 cameras form a circle with two bulges on the left and the right side in the center. Top: view from top, bottom: view from the side.

dows on the facades (Mayer & Reznik 2006).

We have experimented with plane sweeping (Werner & Zisserman 2002), here based on least-squares, to improve the plane parameters derived by RANSAC, but found that for stronger occlusions it is difficult to estimate the bias in brightness. Robust estimation combining, e. g., consensus, with bias determination could be a way to proceed.

We also want to make better use of the information of the planes, e. g., by extending and intersecting planes and checking the newly created planes via homographies, thereby closing gaps. We plan to employ the intersection lines to improve the determination of the vertical direction. Finally, we have started to experiment with the approach by Schnabel et al. (2006), which allows to model point clouds by additional shapes such as cylinders, spheres, and cones.

References

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

Chum, O., Matas, J. & Kittler, J., 2003: Locally Optimized RANSAC. – Pattern Recognition – DAGM 2003, Springer-Verlag, Berlin, Germany, 249–256.

Debevec, P., Taylor, C. & Malik, J., 1996: Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach. – Technical Report CSD-96–893, Computer Science Division, University of California at Berkeley, Berkeley, USA.

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

Fischler, M. & 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): 381–395.

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

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

Mayer, H., 2006: 3D Reconstruction and Visualization of Urban Scenes from Uncalibrated Wide-Baseline Image Sequences. – The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume (36) 5, 207–212.

Mayer, H. & Reznik, S., 2006: MCMC Linked with Implicit Shape Models and Plane Sweeping for 3D Building Facade Interpretation in Image Sequences. – The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume (36) 3, 130–135.

Mikhail, E., Bethel, J. & McGlone, J., 2001: Introduction to Modern Photogrammetry. – John Wiley & Sons, Inc, New York, USA.

Nistér, D., 2004: An Efficient Solution to the Five-Point Relative Pose Problem. – IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (6): 756–770.

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

Schnabel, R., Wahl, R. & Klein, R., 2006: Shape detection in point clouds, Technical Report CG-2006-2, Universität Bonn.

Torr, P., 1997: An Assessment of Information Criteria for Motion Model Selection. – Computer Vision and Pattern Recognition, 47–53.

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

Address of the author:

Prof. Dr.-Ing. Helmut Mayer, Universität der Bundeswehr München, Institut für Photogrammetrie und Kartographie, D-85577 Neubiberg, Tel.: +49-89-6004-3429, Fax: +49-89-6004-4090, e-mail: Helmut.Mayer@unibw.de

Manuskript eingereicht: Januar 2007
Angenommen: Januar 2007