Available online at www.sciencedirect.com

ScienceDirect

ISPRS Journal of Photogrammetry & Remote Sensing 61 (2007) 371–380

PHOTOGRAMMETRY & REMOTE SENSING

www.elsevier.com/locate/isprsjprs

Building facade interpretation from uncalibrated wide-baseline image sequences[☆]

Helmut Mayer [*], Sergiy Reznik

Institute of Photogrammetry and Cartography, Bundeswehr University Munich, D-85577 Neubiberg, Germany

Received 13 October 2005; received in revised form 18 October 2006; accepted 23 October 2006
Available online 15 December 2006

Abstract

We propose an approach for building facade interpretation ranging from uncalibrated wide-baseline image sequences to the extraction of windows. The approach comprises several novel features, such as determination of the facade planes by robust least squares matching, learning of implicit shape models for objects, particularly windows, and the determination of the latter by means of Markov Chain Monte Carlo (MCMC) employing an abstraction hierarchy generated via mathematical morphology. Results for the fully automatic approach show its potential and shortcomings.

Keywords: Facade interpretation; Implicit shape models; Markov Chain Monte Carlo; Abstraction hierarchy; 3D reconstruction

1. Introduction

Automatic interpretation of buildings and particularly their facades gained some interest recently. This is illustrated, for instance 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 facades and together with aerial images and laser-scanner data models of cities. Photogrammetrically inspired work focuses on semi-automatic reconstruction (van den Heuvel, 2001), texturing (Böhm, 2004), and disparity estimation (von Hansen et al., 2004) for facades. 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 facade interpretation from wide-baseline image sequences, especially the extraction of objects such as windows. Concerning the detection of facade 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 facade. Although this works well for textured facades, we found it to be unreliable if the texture is weak.

[☆] Revised and extended version of a paper presented at the joint ISPRS / DAGM workshop 'Object Extraction for 3D City Models, Road Databases and Traffic Monitoring — Concepts, Algorithms, and Evaluation', Vienna, Austria, 29–30 August 2005.

[*] Corresponding author. Tel.: +49 89 6004 3429; fax: +49 89 6004 4090.
E-mail addresses: Helmut.Mayer@unibw.de (H. Mayer), Sergiy.Reznik@unibw.de (S. Reznik).
URL's: http://www.unibw.de/ipk (H. Mayer), http://www.unibw.de/ipk (S. Reznik).

0924-2716/$ - see front matter © 2006 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.isprsjprs.2006.10.007

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 facades. 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 facades. 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 facades.

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 facades and the windows, changing parameters, e.g., their width, brightness, etc., in a way generating 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 a changing number of objects during processing. We also integrate prior as well as image information. We started not changing the number of object instances, therefore, we initially employed 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 facade 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 detect rather reliably 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 geometric constraints based on the fundamental matrix as well as the 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 for all color bands, we sub-pixel precisely determine the 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 F for all consecutive pairs. The epipolar lines derived from 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 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 of about 400 × 400 pixels.

Each triplet linked by $\mathcal{T}$ has its own 3D projective coordinate system. To link the triplets, we use the 3D projective transformation between the last two images of the sequence and the first two images of the current triplet. Additionally, we determine $(n+1)$-fold points by projecting points from the sequence into the third image of the current triplet via $\mathcal{T}$ and we integrate points into the solution, which could not be seen in preceding triplets. When all triplets have been linked on the second or third highest pyramid level, we track all points through the pyramid by 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 by Pollefeys et al. (2004) based on the image of the absolute dual quadric Ω^*. For the latter holds $\omega^* \sim \mathsf{P}\Omega^*\mathsf{P}^\top$, with P the projection matrix for the *i*th camera and the dual image of the absolute conic $\omega^* \sim \mathsf{K}\mathsf{K}^\top$, 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 in the linear computations in the form of prior knowledge on the internal parameters, such as, that the (normalized) principal distance often 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 threefold and 370 fourfold points can be seen on the bottom of Fig. 1, showing on the top four images from Prague's famous Hradschin. The right angle at the building corner has been reconstructed rather well.

3. Determination of facade-planes and-points

Before generating facade planes, we take into account one of the most general constraints for facades, namely being oriented vertically. Using this constraint, we can later safely assume in most cases, that windows or doors are rectangles oriented in parallel to the coordinate axes (x-axis is horizontal and y-axis vertical). 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 up-right or rotated 90°, in the y- or in the x-direction. As it is difficult to decide from the image

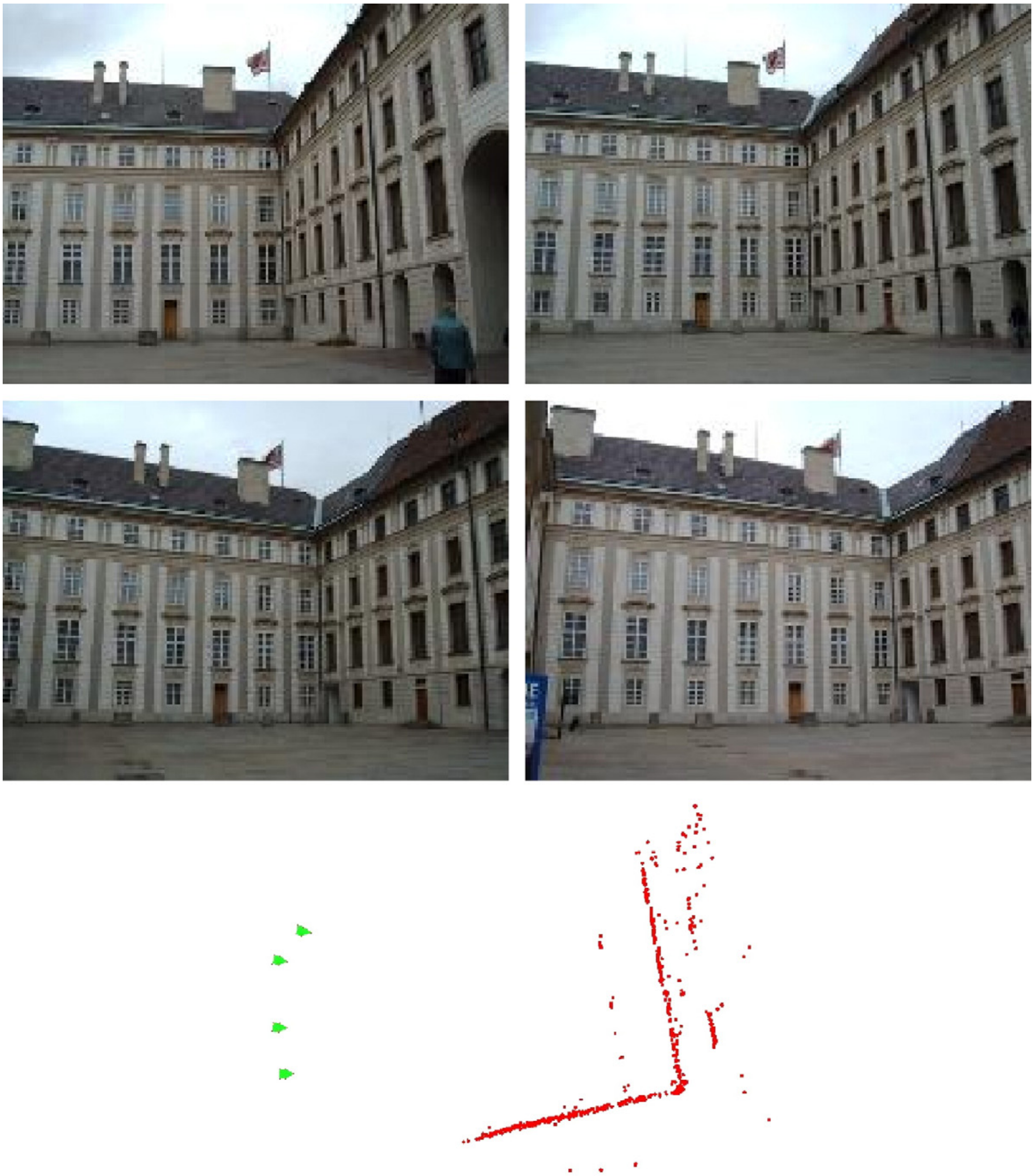

Fig. 1. Wide-baseline quadruple "Prague" (top) and result after orientation and calibration (bottom — points in red, cameras as green pyramids; average standard deviation $\hat{\sigma}_0$=0.24 pixels, image size 4 Mega-pixels).

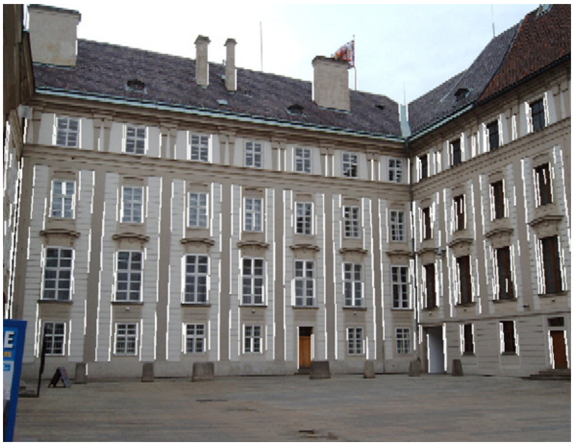

Fig. 2. Lines (white) defining the vertical vanishing points.

alone, in which of these two directions the vertical vanishing point actually lies, we input this information by means of a ?ag, telling that the vanishing point is more in *y*- (standard) or *x*-direction. Everything else is done automatically.

We start by extracting straight lines (Burns et al., 1986). 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 Fig. 2. Knowing the vanishing point and the calibration matrix 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.

Hypotheses for facades are generated in the form of planes from the (Euclidean) 3D points generated as a result of the preceding section. To find 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 this parameter 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 facade 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 facade 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 allow us to transform the information supposed to be on the given plane from all images into the same image geometry.

Fig. 3 shows on the left three of the four images of the Prague scene projected in black-and-white onto one of the facade planes, mapped into the red, green, and blue channel. 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 or in radiometry. This fact is employed by robust least squares optimization of the three plane parameters including the elimination of outlier pixels, implicitly classifying the points on the plane. An alternative would have been to employ Brunn et al. (1996). The result for the classification is given on the right side of Fig. 3. The black parts are supposed to lie on the facade plane, while white holes can be seen as hypotheses for windows, doors, or other architectural elements. Please note that there are several concrete blocks a few meters in front of the facade. This is the reason for the holes at the bottom. Fig. 4 shows both dominant planes computed from about 270 and 250 supporting points, respectively, including the holes and the 3D points.

4. Window detection based on an implicit shape model

As can be seen from Fig. 3 right, the detection of holes in the regions corresponding to a facade plane comprises one possible means to hypothesize windows. Yet, it is not reliable, as windows tend to be dark with low contrast, therefore, not generating outliers, i.e., holes. Thus, 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, an object is modeled 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. For more complex

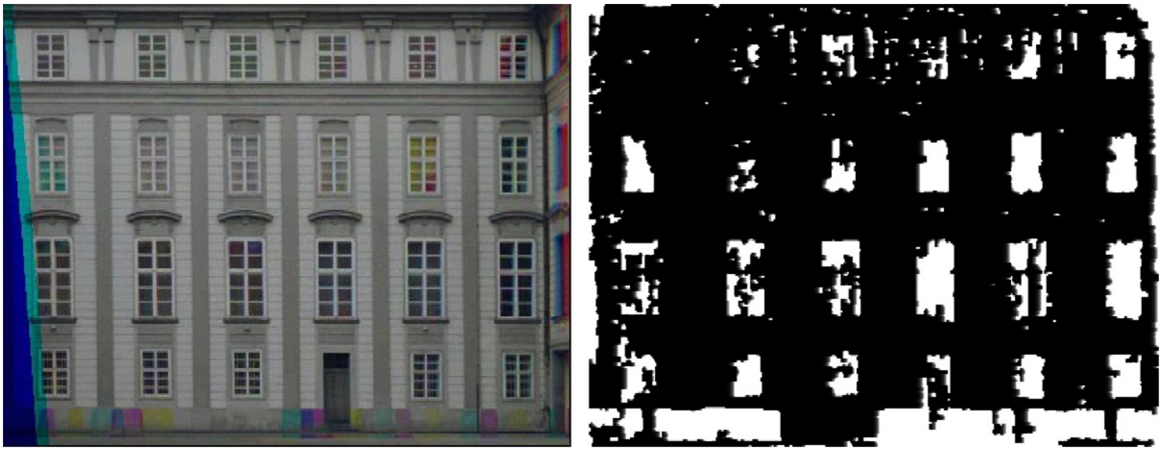

Fig. 3. Three black-and-white images projected onto one of the facade planes mapped into the red, green, and blue image channel showing radiometric differences, but also deviations from the plane (left) and pixels on the plane (black), i.e., pixels not classified as outliers (right).

settings, the SIFT features of Lowe (2004), which are rotation and scale in-variant, are an alternative. 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 facade plane, are oriented by knowing the direction of the vertical vanishing point, and have been scaled to approximately the same pixel size (± 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 (cf. Fig. 5, left). Ten out of 72 windows used for training resulting into 702 points are given in Fig. 5, right.of about 100 × 100 pixels.

To detect windows on a facade, 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 facade as well as the image

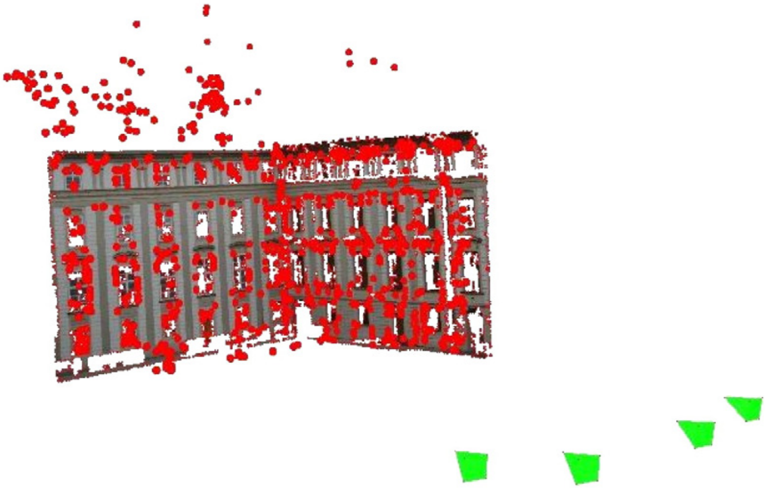

Fig. 4. The two dominant planes restricted to the areas classified to be on the plane together with all 3D points (red) and the cameras (green pyramids).

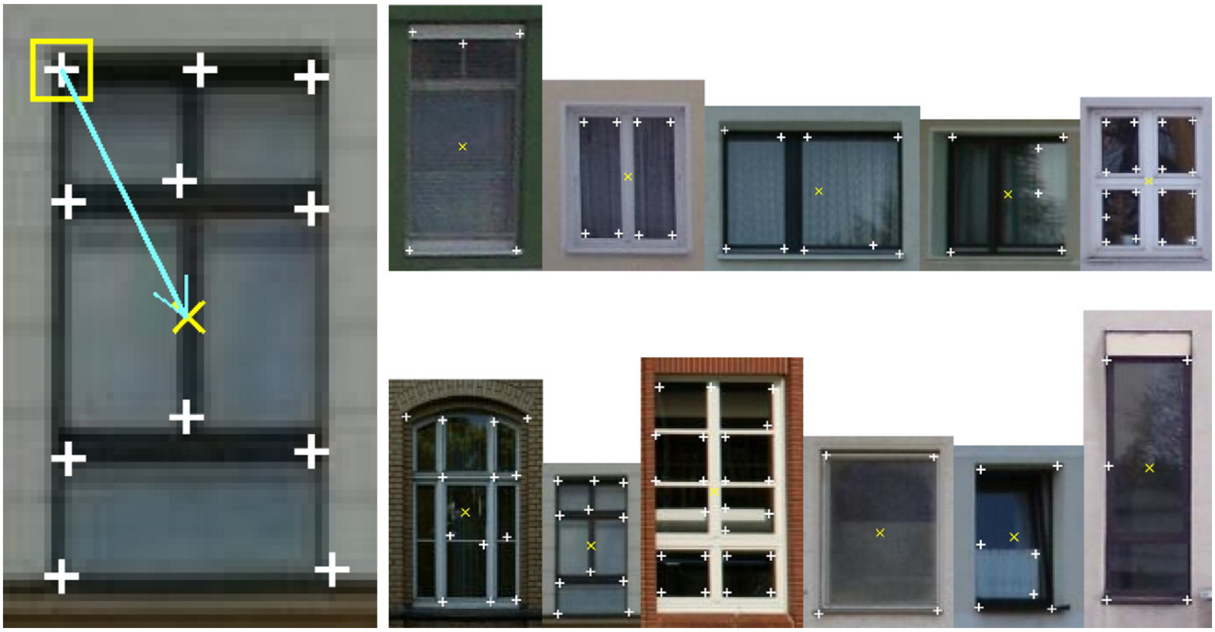

Fig. 5. Left: Example window with Förstner points (horizontal crosses), window center (yellow diagonal cross), one image patch (yellow box), and difference vector to window center (light blue) — right: ten out of 72 windows used for training.

array with the evidence for the position of the window centers are given for our running example in Fig. 6.

Fig. 6 right shows, that the evidence for the window centers is widely spread, because 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 in Fig. 7, showing reasonable hypotheses. Please note, that none of the windows used for training stems from this scene.

5. MCMC based window extraction

Our extraction scheme to determine 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

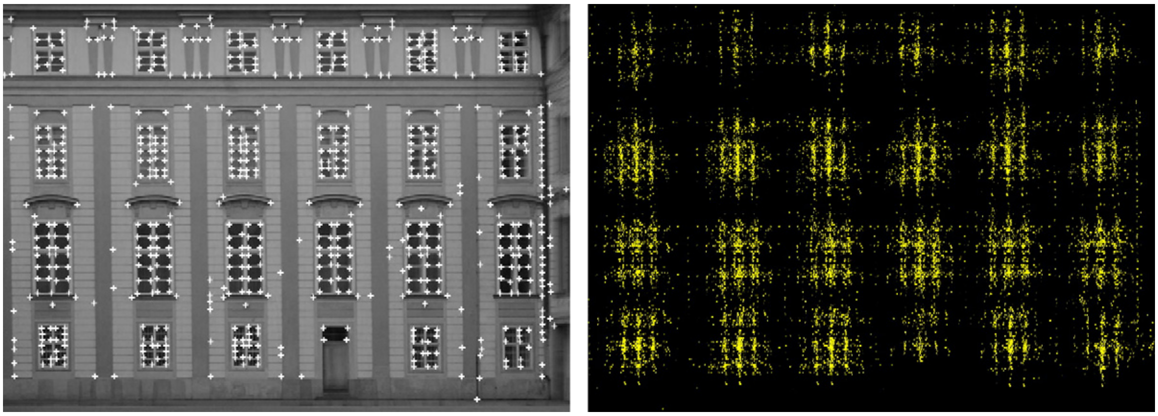

Fig. 6. Facade with Förstner points (crosses — left) and accumulated evidence for centers of windows (right).

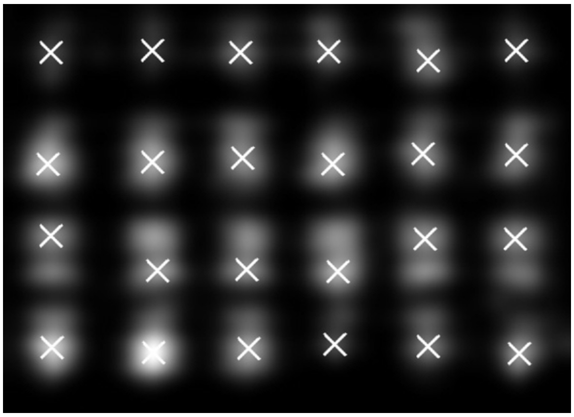

Fig. 7. Accumulated evidence for window centers integrated with Gaussian and maxima (crosses).

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 window boxes or flower pots.

The last experience can be modeled in terms of abstraction by means of a scale-space. What we are interested in at the moment is an object in the range of about 1 × 1.5 m 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 has proven to give particularly useful results for this kind of problem, where objects have a strong 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. Fig. 8, center). The opening before the closing avoids, that bright parts cannot be removed 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

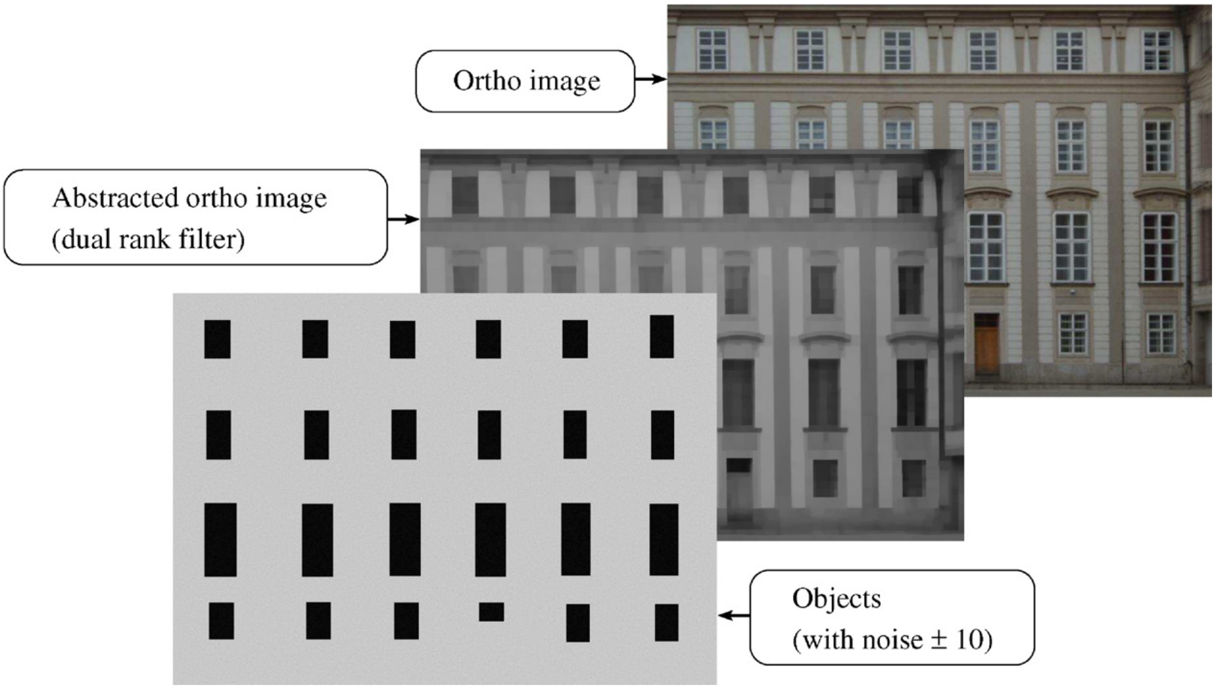

Fig. 8. Abstraction hierarchy consisting of the original image (right), the Dual Rank filtered image (center), and the MCMC model (left).

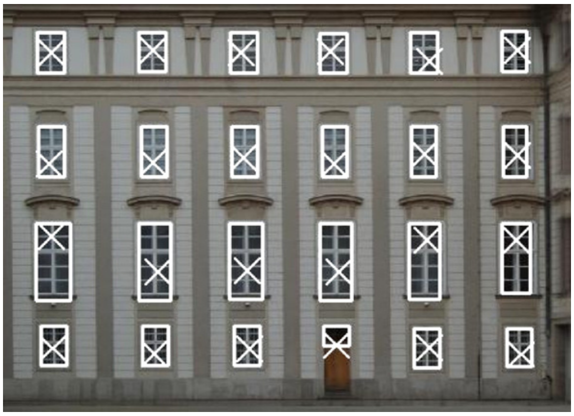

Fig. 9. Result of MCMC — hypotheses (crosses) and windows (boxes).

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. Fig. 8).

The model is disturbed by Gaussian noise and compared to the abstracted facade 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 hypotheses stemming from a procedure determining only the centers of windows, 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 exponentially 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 facade with a window, but only a rectangular image part five times larger than the average window size.

Fig. 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 Fig. 10. For both there is room for improvement for the delineation of the windows, yet all windows have been detected.

Recently, we have started to make use of the fact, that facades often consist of regular structures in the form of rows and/or columns. We model this by an abstraction hierarchy. For it, we employ RJMCMC to in- and exclude objects in the statistical process. Until now, we only generate and eliminate rows of windows, with the probability for the generation as well as for the elimination of a row hypothesis both set to 0.01. A result for this is given in Fig. 11. The left side shows rows of windows with the same size and the same distance

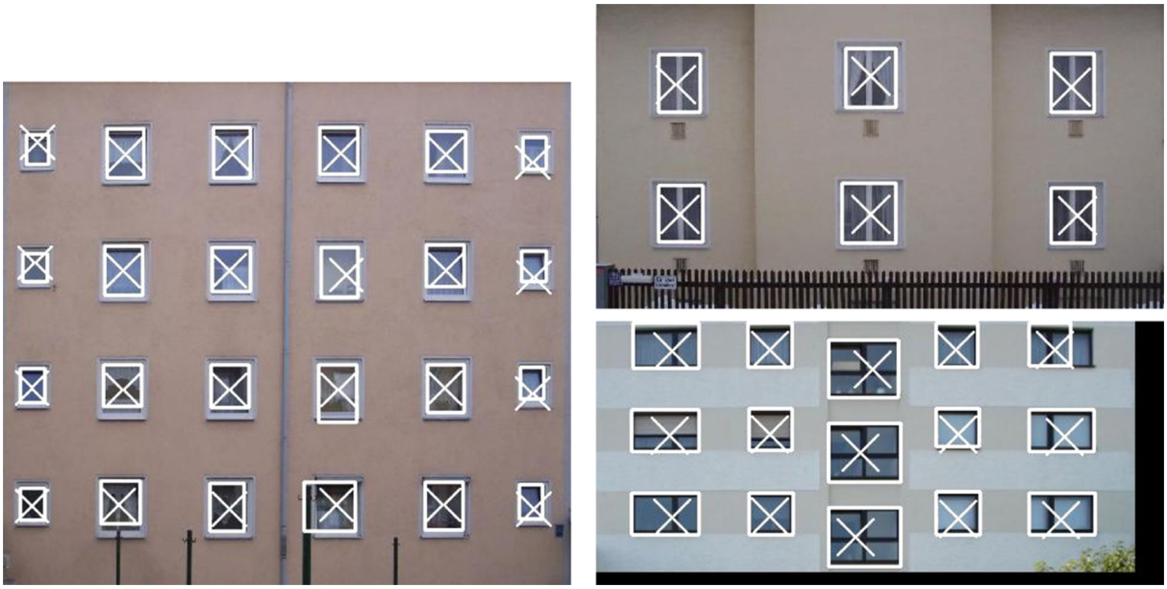

Fig. 10. Further results — hypotheses (crosses) and windows (boxes).

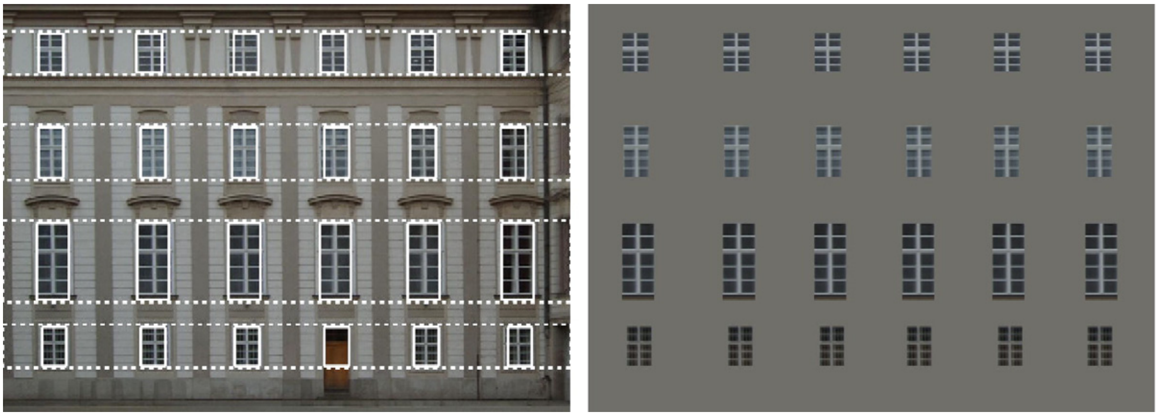

Fig. 11. Rows (dashed boxes) of windows (boxes — left) and visualization based on one prototype per row (right).

between them. Finally, on the right side we demonstrate, how visualization could be made more efficient for larger distances of the viewer by just visualizing prototypes, i.e., the average windows, instead of the individual textures.

6. Conclusions

The results we have presented have been produced fully automatically, using only very few semantically meaningful thresholds, such as for 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 are focusing on the integration of segmentations generated by 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) and Leibe and Schiele (2004). We will also integrate further architectural objects such as doors or columns and we will try to make use of the regular structure of many facades in the spirit of Alegre and Dallaert (2004) based on stochastic, context free grammars. Finally, on a wider time scale, we want to model facades in 3D, by matching the obtained window hypotheses in the images, e.g., by sweeping planes (Werner and Zisserman, 2002), but also by including prominent 3D objects such as balconies.

Acknowledgments

Sergiy Reznik is funded by Deutsche Forschungsgemeinschaft under grant MA1651/10. We thank the anonymous reviewers for their helpful comments.

References

Agarwal, S., Awan, A., 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), 1475–1490.

Alegre, F., Dallaert, F., 2004. A probabilistic approach to the semantic interpretation of building facades. International Workshop on Vision Techniques Applied to the Rehabilitation of City Centres. 12 pages.

Bauer, J., Karner, K., Schindler, K., Klaus, A., Zach, C., 2003. Segmentation of building models from dense 3d point-clouds. 27th Workshop of the Austrian Association for Pattern Recognition.

Böhm, J., 2004. Multi image fusion for occlusion-free façade texturing. The International Archives of the Photogrammetry. Remote Sensing and Spatial Information Sciences, vol. 35 (Part B5), pp. 867–872.

Brunn, A., Lang, F., Förstner, W., 1996. A procedure for segmenting surfaces by symbolic and iconic image fusion. Mustererkennung 1996. Springer-Verlag, Berlin, Germany, pp. 11–20.

Burns, J., Hanson, A., Riseman, E., 1986. Extracting straight lines. IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (4), 425–455.

Dick, A., Torr, P., Cipolla, R., 2004. Modelling and interpretation of architecture from several images. International Journal of Computer Vision 60 (2), 111–134.

Eckstein, W., Munkelt, O., 1995. Extracting objects from digital terrain models. Remote Sensing and Reconstruction for Three-Dimensional Objects and Scenes. SPIE, vol. 2572, pp. 43–51.

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, pp. 281–305.

Früh, C., Zakhor, A., 2003. Constructing 3d city models by merging aerial and ground views. IEEE Computer Graphics and Applications 23 (6), 52–61.

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

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

Leibe, B., Schiele, B., 2004. Scale-invariant object categorization using a scale-adaptive mean-shift search. Pattern Recognition - DAGM 2004. Springer-Verlag, Berlin, Germany, pp. 145–153.

Lowe, D., 2004. Distintive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60 (2), 91–110.

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., Tops, J., 2004. Visual modeling with a hand-held camera. International Journal of Computer Vision 59 (3), 207–232.

Ribarsky, A., Rushmeier, H., 2003. 3D reconstruction and visualization. IEEE Computer Graphics and Applications 23 (6), 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, 247–260.

von Hansen, W., Thonnessen, U., Stilla, U., 2004. Detailed relief modeling of building facades from video sequences. The International Archives of the Photogrammetry. Remote Sensing and Spatial Information Sciences, vol. 35 (Part B3), pp. 967–972.

Wang, X., Totaro, S., Taillandier, F., Hanson, A., Teller, S., 2002. Recovering facade texture and microstructure from real-world images. The International Archives of the Photogrammetry. Remote Sensing and Spatial Information Sciences, vol. 34 (Part 3A), pp. 381–386.

Werner, T., Zisserman, A., 2002. New techniques for automated architectural reconstruction from photographs. Seventh European Conference on Computer Vision, vol. II, pp. 541–555.

Wilczkowiak, M., Sturm, P., Boyer, E., 2005. Using geometric constraints through parallelepipeds for calibration and 3d modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2), 194–207.