MCMC LINKED WITH IMPLICIT SHAPE MODELS AND PLANE SWEEPING FOR 3D BUILDING FACADE INTERPRETATION IN IMAGE SEQUENCES

Helmut Mayer and Sergiy Reznik

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

KEY WORDS: Markov Chain Monte Carlo, Implicit Shape Models, Plane Sweeping, Facade Interpretation

ABSTRACT:

In this paper we propose to link Markov Chain Monte Carlo – MCMC in the spirit of (Dick, Torr, and Cipolla, 2004) with information from Implicit Shape Models – ISM (Leibe and Schiele, 2004) and with Plane Sweeping (Werner and Zisserman, 2002) for the 3D interpretation of building facades, particularly for determining windows and their 3D extent. The approach starts with a (possibly uncalibrated) image sequence, from which the 3D structure and especially the vertical facades are determined. Windows are then detected via ISM. The main novelty of our work lies in using the information learned by the ISM also to delineate the window extent. Additionally, we determine the 3D position of the windows by plane sweeping in multiple images. Results show potentials and problems of the proposed approach.

1. INTRODUCTION

Recently, there are – among others – two yet not contradicting important directions in object extraction: Appearance based and generative models. Prominent examples for the former are, e.g., (Lowe, 2004) and (Agarwal, Awan, and Roth, 2004). The basic idea of these two and similar approaches is that an object is modeled by features computed from small characteristic image patches and their spatial arrangement, both being learned more or less automatically from given training data, i.e., images. While this can also be seen as a discriminative model where a hypothesis for an object is created bottom-up from the data, generative models go the other way, i.e., top-down: From a given hypothesis they generate a plausible instance of the data generatively, i.e., via computer graphics, and compare it with the given image data. Usually this is done in a Bayesian framework. There are priors for the parameters, the comparison with the data results into a likelihood, and both are combined into the posterior. One particularly impressive example for an approach linking discriminative and generative modeling tightly in a statistically sound manner is (Tu, Chen, Yuille, and Zhu, 2005). In (Fei-Fei, Fergus, and Perona, 2004) a generative appearance based model is employed to learn 101 object categories from only a few training examples for each class via incremental Bayesian learning.

We are aiming at the interpretation of building facades from image sequences, particularly inspired by the generative model based on Markov Chain Monte Carlo – MCMC, e.g., (Neal, 1993), put forward in (Dick, Torr, and Cipolla, 2004). To detect objects, in our case windows, we follow (Mayer and Reznik, 2005) who use appearance based modeling in the form of an Implicit Shape Model – ISM, as introduced by (Leibe and Schiele, 2004). Yet, and this is the main novelty of our approach, we additionally link ISM to MCMC for the determination of the window extent. By this means we partly avoid the tedious manual generation of a model for the in our case sometimes complex structures of windows and also robustify the approach. Additionally, we compute the three-dimensional (3D) extent of the windows by means of plane sweeping proposed in (Werner and Zisserman, 2002). Opposed to (Werner and Zisserman, 2002) as well as (Bauer, Karner, Schindler, Klaus, and Zach, 2003) we do not detect windows as objects which are situated behind the facade plane which makes us independent from the fact if the windows are behind, on, or in even in front of the facade.

The basic idea of the generative model of (Dick, Torr, and Cipolla, 2004), which is our main inspiration, is to construct the building from parts, such as the facades and the windows, for which parameters, e.g., the width, brightness, are changed statistically to produce an appearance resembling the images after perspectively projecting the model with the given parameters. The difference between the given and the generated image determines the likelihood that the data fits to the model and is combined with prior information describing typical characteristics of buildings.

Other work on facades is, e.g., (Früh and Zakhor, 2003), where a laser-scanner and a camera mounted on a car are employed to generate 3D models of facades (yet without information about objects such as windows or doors) and together with aerial images and aerial laser-scanner data realistic models of areas of cities. In photogrammetry as well as in computer vision semi-automatic approaches have been proposed (van den Heuvel, 2001; Wilczkowiak, Sturm, and Boyer, 2005), where the latter exploits special geometrical constraints of buildings for camera calibration. (Böhm, 2004) shows how to eliminate visual artifacts from facades by mapping images from different view points on the facade plane employing the robust median. The determination of fine 3D structure on facades via disparity estimation is presented by (von Hansen, Thönnessen, and Stilla, 2004). (Wang, Totaro, Taillandier, Hanson, and Teller, 2002) take into account the grid, i.e., row / column, structure of the windows on many facades. (Alegre and Dallaert, 2004) propose a more sophisticated approach, where a stochastic context-free grammar is employed to represent recursive regular structures of the windows. Both papers only give results for one or two very regular high-rising buildings.

In Section 2. we sketch our approach to generate a vertically oriented Euclidean 3D model consisting of cameras and points from (possibly uncalibrated) image sequences, from which we determine vertical facade planes. Section 3. describes the ISM and as main contribution of this paper how we learn and use the segmentation information to help delineate the windows via MCMC. Finally, in Section 4. we show how the 3D extent of the windows can be determined based on plane sweeping. The paper ends up with conclusions.

2. 3D RECONSTRUCTION

Our approach is based on wide-baseline image sequences. After projective reconstruction using fundamental matrices and trifocal

tensors (Hartley and Zisserman, 2003) employing Random Sample Consensus – RANSAC (Fischler and Bolles, 1981) based on Förstner points (Förstner and Gülch, 1987) which we match via least squares matching, we calibrate the camera employing the approach proposed in (Pollefeys, Van Gool, Vergauwen, Verbiest, Cornelis, and Tops, 2004). If calibration information is available, we use (Nistér, 2004) to determine the Euclidean 3D structure for image pairs. Our approach deals efficiently with large images by using image pyramids and we obtain full covariance matrices for the projection matrices and the 3D points by means of bundle adjustment taking into account the covariance matrices of the least squares matching of all employed images.

Having generated a 3D Euclidean model we orient it vertically based on the vertical vanishing point derived from the vertical lines on the facade and the given calibration parameters. The vertical vanishing point is detected robustly again using RANSAC, the user only providing the information if the camera has been been very approximately held horizontally or vertically.

The vertically oriented model is the basis for the determination of the facade planes using once again RANSAC. To make the determination more robust and precise, we employ the covariance information of the 3D points from the bundle adjustment by testing the distances to a hypothesized plane based on the geometric robust information criterion – GRIC (Torr, 1997). Additionally, we check, if the planes are vertical and we allow only a limited overlap of about five percent between the planes. The latter is needed, because of the points possibly situated on intersection lines between the planes.

Finally, as the position of the facade planes is often determined in-between the plane defined by the real facade and the plane defined by the windows, its depth is optimized via plane sweeping (Baillard and Zisserman, 1999; Werner and Zisserman, 2002). From the parameters for the facade planes as well as the projection matrices we compute homographies between the plane and the images. We project all images a facade can be seen from (this can be derived via the points that lead to the plane and from which images they were determined) onto the facade plane and compute an average image as well as the bias in brightness for each projected image to it. Then, we move the facade plane in its normal direction and determine for a larger number of distances the squared differences of gray values to the average image for all images after subtracting the bias in brightness determined above. We finally take the position, where this difference is minimum.

The result of this step are projection matrices, 3D points, and optimized facade planes all in a vertically oriented Euclidean system. The only additional information the user has to provide for the further processing is the approximate scaling of the model so that the images can be projected on the facade with a normalized pixel-size. Therefore, for the next step of the delineation of windows on the facade we can assume vertically oriented facade planes with a standardized pixel size.

3. DETECTION AND DELINEATION OF WINDOWS BASED ON MCMC AND ISM

An Implicit Shape Model – ISM (Leibe and Schiele, 2004) describes an object in the form of the spatial arrangement of characteristic parts. As (Agarwal, Awan, and Roth, 2004) we use as parts image patches (here of the empirically determined size 9×9 pixels) around Förstner points. Training patches and patches in an image to be analyzed are compared via the (normalized) cross correlation coefficient (CCC). For the arrangement of the points we employ as (Leibe and Schiele, 2004) the generalized Hough transform.

Similarly as (Mayer and Reznik, 2005), we "learn" the model for a window in a way that can be seen as a simplified version of (Leibe and Schiele, 2004): We manually cut out image parts containing training windows using in the range of about 100 windows. In these (cf. Figure 1 for an example) we extract Förstner points with a fixed set of parameters. Opposed to (Mayer and Reznik, 2005), we manually mark the extent of the whole window including the frame and compute from it the center.

We "learn" only salient points at the corners of the manually marked window extent (small yellow squares in Figure 1). For these we store the gray values in the patches around the points, their relation to the window center in the form of the difference vector, and particularly their relation to the window extent. This is done in the form of images of the edges of the window extent. The latter gives information which we use for the segmentation, i.e., the delineation of the window, the main novelty of our approach. Figure 2 shows examples for image patches (left) together with the edges derived from the manually given window extent (right). Please note that for many of our (training) windows the window extent does not fit too well to the Förstner points as they tend to be situated at the salient image corner between glass and window frame.

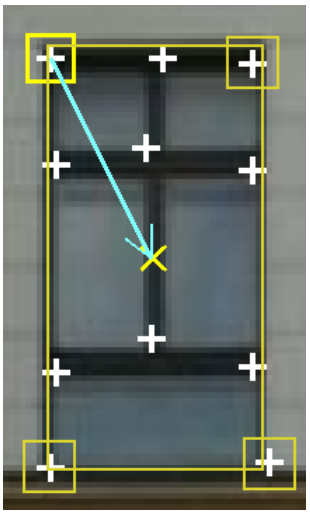


Figure 1. Image part containing training window with Förstner points (white crosses), manually marked window extent (yellow rectangle), window center (yellow diagonal cross), patches around salient points at the corners of the window extent (small yellow squares), and one of four difference vectors to center (blue arrow)

To detect windows on a facade, we extract Förstner points with the same set of parameters as above (cf., e.g., Figure 3, left) and compare the patches of size 9×9 centered at them with all salient points learned above by means of CCC. If CCC is above an empirically determined threshold of 0.9, we write out the difference vector learned for the corresponding point into an initially empty evidence image, incrementing the corresponding pixel by one. By this means, each match votes for the position of the window center. The Förstner points as well as the evidence for the position of the window centers are given for our running example in Figure 3, right.

Figure 3, right, shows that the hypothesized window centers are widely spread, because parts of windows can vote for different

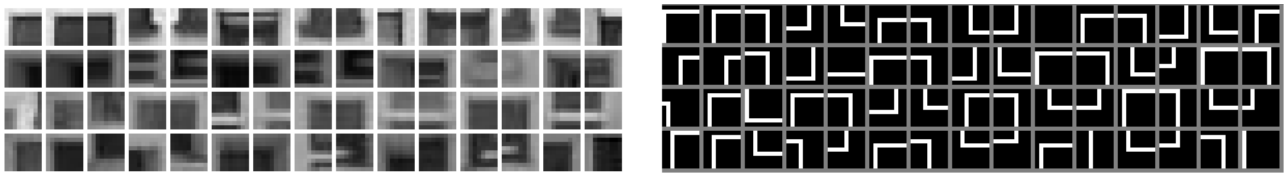


Figure 2. Set of patches (left) and set of edges (right) for window corners

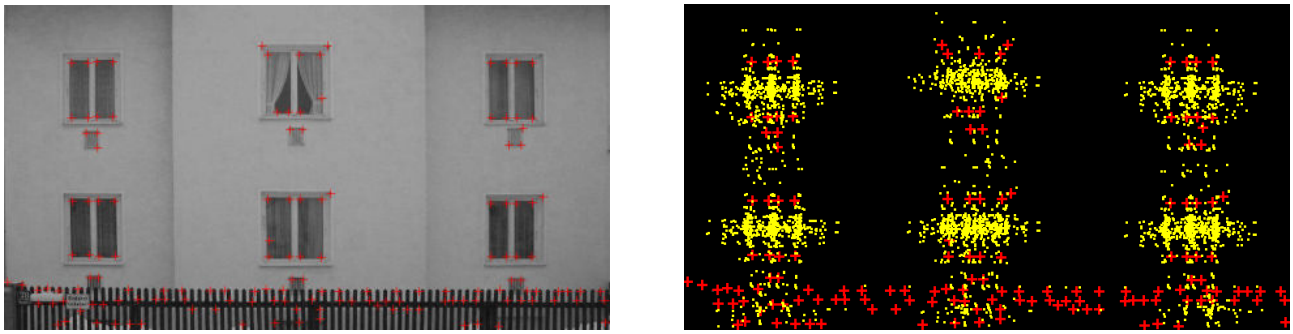


Figure 3. Facade (left) and evidence for window centers (yellow dots, right) both with Förstner points (red crosses)

positions. A patch can look, e.g., similar to an upper right corner of a whole window, but is actually situated at a transom (horizontal bar) at the center of the window. To generate meaningful hypotheses for window centers, we, therefore, integrate the evidence by smoothing them with a Gaussian and then determine all local maxima above a given threshold. The result for this is shown in Figure 5, left. Please note that none of the windows used for training stems from this scene as well as any of our examples presented in this paper.

The information from the ISM is used for segmentation by inserting it into the generative modeling based on MCMC. For this, the patches voting for the respective centers need to be determined. In Figure 5, right, all hypotheses and their difference vectors for the areas around the local maxima for the window centers, where the evidence is beyond 0.9 of the local maximum value, are shown. From these vectors only the vectors pointing diagonally are retained. Only they provide information about the window extent, because windows are assumed not to be extremely narrow or low. The average vectors of these patches pointing to the center are shown in Figure 4 together with the areas where the evidence is locally above 0.9 of its maximum.

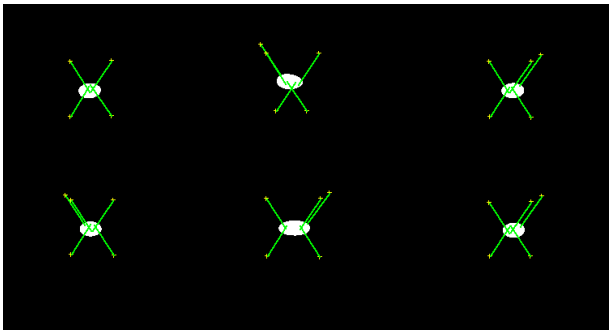


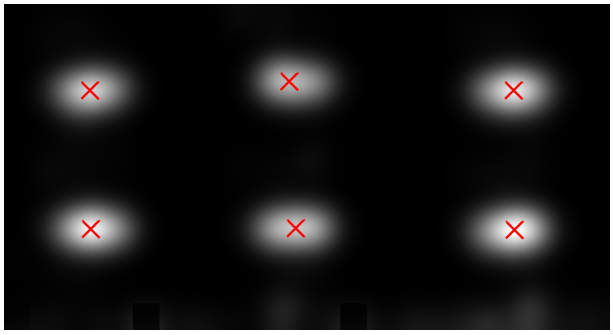
Figure 4. Areas with a value beyond 0.9 times of the local maxima (white) and average vectors of all hypotheses for corners pointing diagonally to the maxima, i.e., hypotheses for the window centers (green lines)

Once the potential patches at the window corners are known, the corresponding edges (cf. Figure 2, right) are summed up (cf. Figure 6, left). For guiding MCMC, the edges are thinned, normalized and then blurred to extend the area of convergence. As the likelihood is normalized in the MCMC process, it is important that the ends of the straight segments are cut and not blurred in the direction of the edge. The result is the window corner image (cf. Figure 6, right).

To delineate the windows, we start with hypotheses constructed from the centers of the diagonally most distant patches voting for a particular window and a small inward offset of 8 pixels in horizontal and vertical direction to avoid that the random search starts outside the window extent. We then take up the basic idea of (Dick, Torr, and Cipolla, 2004), i.e., we try to generate an image which is similar to the actual image. Our basic model is very simple, namely a rectangle brighter or darker than the background, i.e., with an edge to the background. The corresponding edges for the windows are projected into the window corner image and the normalized strength of all pixels above zero gives the likelihood. As we found that for bright facades it is very helpful that windows are in most cases darker than the facade plane, we follow for them (Mayer and Reznik, 2005) and correlate a model consisting of noisy dark rectangles on a bright background with the facade image abstracted by gray-scale morphology. The result for this is then combined with the result based on ISM on a half and half basis.

Figure 7, left, shows a hypothesis for the window extent, i.e., the start position, and right the final position. Please note that we have employed the half and half combination of correlation and ISM for the running example with its bright facade. Therefore, the final position in Figure 7, right, does not fit perfectly to the distribution given by the ISM.

The parameters for the window extent are disturbed by Gaussian noise taking into account the prior that the ratio of height to width of a window lies in the majority of cases between 0.25 to 5 modeled by a mixture of Gaussians. For each iteration of MCMC, we either change the width, the height, or the position of the rectangle representing the window extent. For robustification we use simulated annealing. I.e., the higher the number of iteration becomes, the lower becomes the probability to accept results which



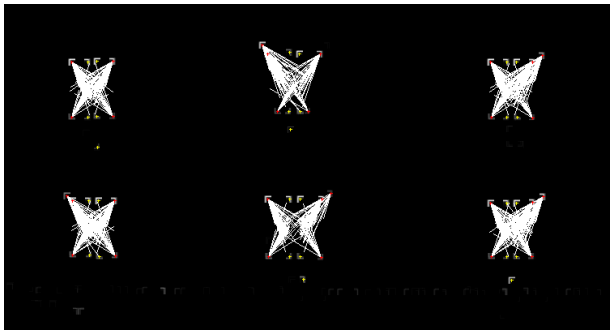


Figure 5. Evidence for window centers integrated with Gaussian together with maxima (diagonal red crosses – left) and hypotheses for window corners pointing to the maxima (right)

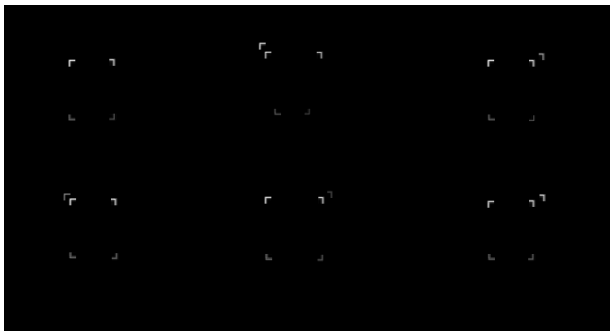
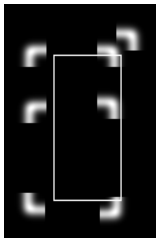


Figure 6. Sum of edges describing window corners (left) and derived distribution to guide MCMC (window corner image, right)



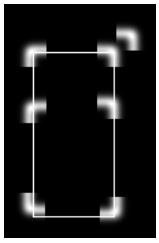


Figure 7. Distribution from ISM used to guide MCMC with hypothesis for window extent, i.e., start position (left) and final position (right).

are worse than for the preceding iteration. Figure 8 shows the hypotheses in white and the final result in green.

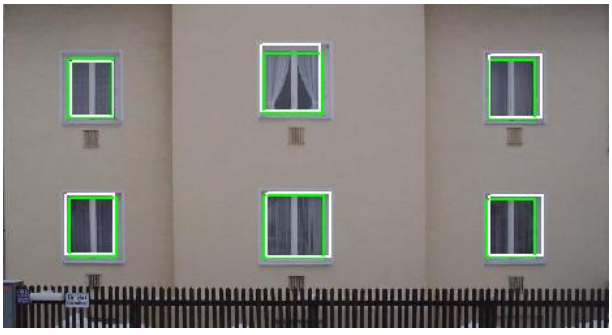


Figure 8. Hypotheses for the window extent (white) and final outcome (green)

4. DETERMINATION OF THE 3D EXTENT OF WINDOWS VIA PLANE SWEEPING

As windows are often not lying on the facade plane, but mostly behind it, their 3D position needs to be determined. This is done again by means of plane sweeping, cf. Section 2., employing the 3D Euclidean reconstruction result by computing homographies between planes and images. The bias in brightness of the images to an average image is computed for the whole facades as it is too unreliable for the individual windows. To determine the depth of a particular window, we move the rectangular part of the facade plane determined above to correspond to a window in the direction of the normal of the facade plane. We compute for a larger number of reasonable distances from the facade plane the squared differences of gray values from the individual images it can be seen from to the average image and take the position, where the difference is minimum.

Results for this are given in Figures 10 and 13. The first result, the input images for which are given in Figure 9, shows that we are actually dealing with a 3D setup where not only images of facade planes, but also there 3D position and relations to the cameras are known. For this bright facade again ISM and correlation have been used on a half by half basis leading to a meaningful delineation of the windows after detecting all windows on the facade. Also plane sweeping was successful for all windows as can be seen from the nearly constant offset. With our approach we are able to determine different depths for individual windows as we do not employ 3D information in the form of local maxima of the whole plane to determine possible window hypotheses such as (Werner and Zisserman, 2002). Yet, we have to note that a combination of both ideas might be the best way to proceed to deal with more complex situations.

For the second building in Figure 13 (input images cf. Figure 12,

Figure 9. Four images used to generate the model given in Figure 10

Figure 10. Bright building seen from the back – the windows are marked in red on the facade and in green behind the facade; cameras are given as green pyramids with the tip defining the projection center and the base the viewing direction

3D points and cameras, cf. Figure 11) the facades are rather dark. Therefore, we could only use ISM for the delineation of the windows. One can see from Figure 13, left, that for it all windows have been detected, except for the upper left, where the resolution of the image is not good and which is disturbed by a bird house. As we have not yet modeled doors, the door on the right facade is interpreted as a window. Figure 13, right, shows that in most cases there was a correct and consistent determination of the depth of the windows. Here one has to note, that these are mostly windows without mullions and transoms, where a determination of the depth is rather difficult, also because the windows are partly reflecting the surroundings.

Figure 11. 3D points and cameras (green pyramids) for dark building

5. CONCLUSIONS

We have shown how by combining appearance based and generative modeling employing MCMC and ISM the extent of objects, particularly windows, can be determined robustly based on automatically learned models even if the structure of the object varies or the contrast is weak. This can be seen as an extension of approaches such as (Dick, Torr, and Cipolla, 2004), where a less adaptive object-wise modeling of the texture was employed. We have also demonstrated how based on plane sweeping employing homographies between the facade plane and the images it is possible to determine the 3D position of the planar hypotheses for the windows.

Windows, but also other objects on the facade, can have substructures of different sizes, e.g., mullions and transoms. To model them and also other objects such as doors and architectural details, we plan to integrate scale into ISM.

The homographies employed in the 3D determination could on one hand help to identify 3D details not lying on the facade, but could also be used to compute the 3D position of (partly) planar objects far off, but parallel to the facade plane such as balconies. For handling problems with different reflectivity we plan to introduce a robust estimator.

To be able to model rows or columns of windows or architectural details and grids made up of them, it is essential that one can deal with models of changing complexity. A means for this is Reversible Jump Markov Chain Monte Carlo – RJMCMC (Green, 1995), used, e.g., by (Dick, Torr, and Cipolla, 2004). It allows to change the number of objects during processing, i.e., to include new windows, etc. To model rows, columns, and grids in a principled way we want to employ a (context free) stochastical grammar describing the hierarchy of the objects on the facade as well as of the different facades of a building in the spirit of (Alegre and Dallaert, 2004).

Figure 12. Four images used to generate the model given in Figures 13 and 11

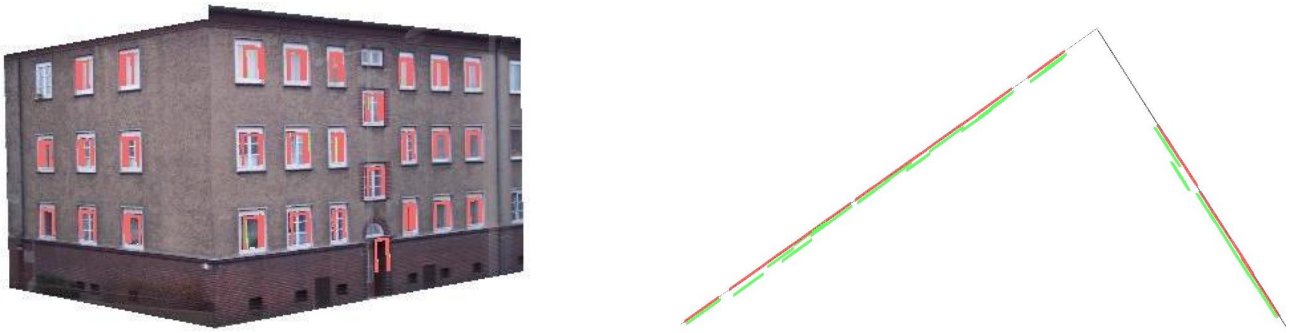


Figure 13. Dark building seen from the outside (left) and from the top (right) – colors and cameras cf. Figure 10

ACKNOWLEDGMENTS

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

Baillard, C. and Zisserman, A., 1999. Automatic Reconstruction of Piecewise Planar Models from Multiple Views. In *Computer Vision and Pattern Recognition*, Volume II, pp. 559–565.

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*, Volume (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), 111–134.

Fei-Fei, L., Fergus, R., and Perona, P., 2004. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In *IEEE Workshop on Generative-Model Based Vision*.

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), 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), 52–61.

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

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

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

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

Mayer, H. and Reznik, S., 2005. Building Façade Interpretation from Image Sequences. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume (36) 3/W24, pp. 55–60.

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

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., and Tops, J., 2004. Visual Modeling with a Hand-Held Camera. *International Journal of Computer Vision* 59(3), 207–232.

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

Tu, Z., Chen, X., Yuille, A., and Zhu, S.-C., 2005. Image Parsing: Unifying Segmentation Detection and Recognition. *International Journal of Computer Vision* 63(2), 113–140.

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., 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*, Volume (35) B3, pp. 967–972.

Wang, X., Totaro, S., Taillandier, F., Hanson, A., and Teller, S., 2002. Recovering Facade Texture and Microstructure from Real-World Images. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume (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*, Volume 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), 194–207.