What Next: Autonomous Photogrammetric Image Understanding?

HELMUT MAYER, Munich

ABSTRACT

This paper tries to take up the challenge issued by the title given by the organizers of the photogrammetric week. Due to the vastness of the problem we do so in a selective manner. We focus on two aspects of image understanding with broad recent interest and bright scientific perspective, but also limitations which need to be overcome. These are particularly appearance based approaches for object extraction and classification as well as statistical and combinatorial modeling. We illustrate the paper with examples of our work on building façade interpretation resting on developments in three-dimensional (3D) reconstruction from uncalibrated image sequences. Particularly, we introduce implicit shape models as well as Random Sample Consensus (RANSAC) and Markov Chain Monte Carlo (MCMC). We finally note, that in spite of the large progress in image understanding in recent years, the gap to what people would like to have still seems to widen

1. INTRODUCTION

The title given by the organizers issues a challenge which we were happy to take up on one hand, but which is also hard to fulfill on the other hand. We therefore decided to limit our scope to two areas with broad recent interest, namely appearance based approaches for object extraction and classification as well as statistical and combinatorial modeling.

Appearance based approaches, i.e., approaches, where image information is directly used to model an object, have been around for a while. Yet, only with recent work they have become a main stream of research focusing on the extraction of objects from images, e.g., (Agarwal et al. 2004, Leibe et al. 2004) and to some extent (Lowe 2004), but also on the classification of whole images. For the latter, Fei-Fei et al. (2004) have shown how to incrementally learn and discriminate 101 object classes. For all the above approaches, the basic idea is to combine the comparison of small patches of the image around salient points with the modeling of the spatial arrangement of these points. The big advantage of doing so is, that the model can be learnt automatically from images or their parts tagged to be examples of a certain class.

Mumford (2000) has declared 'The Dawning of the Age of Stochasticity'. While photogrammetry has always been doing statistics in conjunction with bundle-adjustment, one is usually happy to employ deterministic methods, where the result for given data is always the same. Non-deterministic combinatorial approaches such as random sample consensus, or short RANSAC (Fischler and Bolles, 1981) as well as Markov Chain Monte Carlo (MCMC – Neal 1993) have abandoned determinism. While one might feel not at ease with obtaining a different, and sometimes even wrong result for every new run of an algorithm, it is extremely important to note, what one gains: the ability to solve problems, which could not be solved before, or at least not in a reasonable amount of time.

Before discussing appearance based approaches in Section 3 and statistical and combinatorial modeling in Section 4, we give a short account of our recent developments in three-dimensional (3D) reconstruction from uncalibrated imagery to start our running example of building façade interpretation. The paper ends up with conclusions.

2. 3D RECONSTRUCTION

With Hartley and Zisserman (2003), the second edition of a book initially published in 2000, 3D reconstruction based on projective geometry has become textbook knowledge. It makes it possible to generate a Euclidean 3D model from nothing, but perspective images alone. Euclidean 3D model means in this case a model like in photogrammetric relative orientation: Only location, i.e., x-, y-, and z-coordinates of the points, the three parameters for orientation, and the scale are unknown, but right angles are projected to right angles, and ratios of lengths as well as general angles in space are preserved.

The basic ingredients for projective geometry based 3D reconstruction are the fundamental matrix and the trifocal tensor, linking linearly two or three, respectively, perspective images. Combined with RANSAC, one can obtain a correct solution even for cases where the image matching results into less than 20% correct matches. The basic idea of RANSAC is to compute a large number of solutions from randomly selected points (to keep the complexity down, the minimum number of points necessary for a solution is used) and decide about the best one by measuring the support a solution receives from the remaining points. For the fundamental matrix, the support is, e.g., nothing else but the number of points which are closer than a given threshold, e.g., half a pixel, to their corresponding epipolar lines which can be computed by matrix – vector multiplication from the fundamental matrix. While one usually always gets another result from RANSAC for every run, it still makes it possible to solve a problem, for which standard robust estimation is infeasible.

The triplets are linked into projective sequences and blocks. To obtain a metric solution, auto-calibration is needed. Here, we employ the work of Pollefeys et al. (2004) based on the dual image of the absolute conic. In Figure 1, four of eight images from a backyard in Bonn taken with a consumer Sony P 100 camera with 5 Megapixels and a Zeiss objective as well as the 3D model generated fully automatically are shown. The right angles have been reconstructed very well, there are many n-fold points (with $n \geq 3$) and the backprojection error is 0.3 pixels. By comparison with other image sequences we found that, given a suitable 3D geometry, the camera parameters can be estimated in the range of a few percent. Noting, that automatic 3D reconstruction from scratch starts to become operational, commission III of the International Society for Photogrammetry and Remote

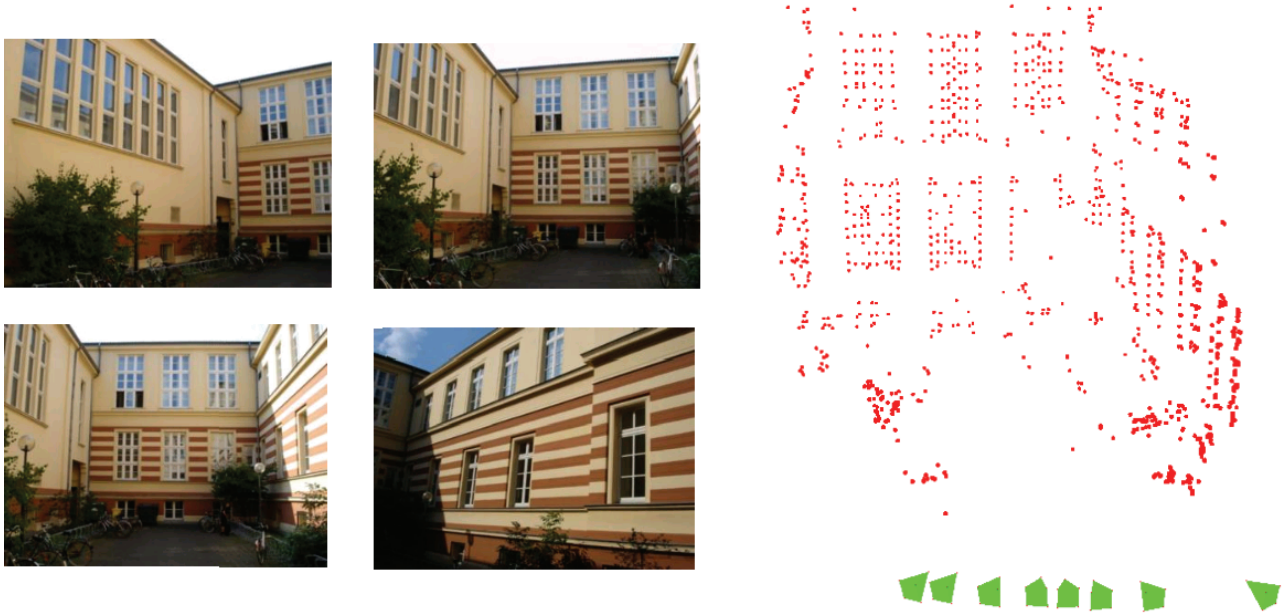

Figure 1: Bonn, backyard, first three and last image as well as 426 3-fold, 377 4-fold, 288 5-fold, 103 6-fold, and 20 7-fold points (red) as well as camera positions (green pyramids), $\sigma_0 = 0.3$ pixels

Sensing (ISPRS) has recently set up a working group with one of the important goals being to demonstrate the potential and the limits of 3D reconstruction by means of an international test in conjunction with the computer vision community.

We use the result of the above 3D reconstruction to generate hypotheses for the objects we are interested in, namely building façades (cf. also Mayer and Reznik 2005). As the vertical direction plays a major role for façades, we first determine it from vertical lines typically found on façades. For the computation of the vertical vanishing point where these lines meet as well as to derive the façade planes from the given 3D points, we employ RANSAC. For the façade planes this means, that we choose three points randomly, construct a plane from them, and check how many of the given 3D points belong to this plane. We finally take those planes with maximum support, but with only a tiny overlap (this happens at the line of intersection).
The images are then projected onto the planes and by means of least squares matching the parameters of the planes are improved while at the same time areas not on the plane are determined by robust estimation. The result for the running example for the rest of the paper from the Hradschin in Prague, Czechia, is given in Figure 2. For the two planes there have been about 270 and 250 supporting points, respectively. Some of the parts not on the façades have been found, but if there is not enough local contrast, robust estimation fails to detect particularly the windows behind the façades.

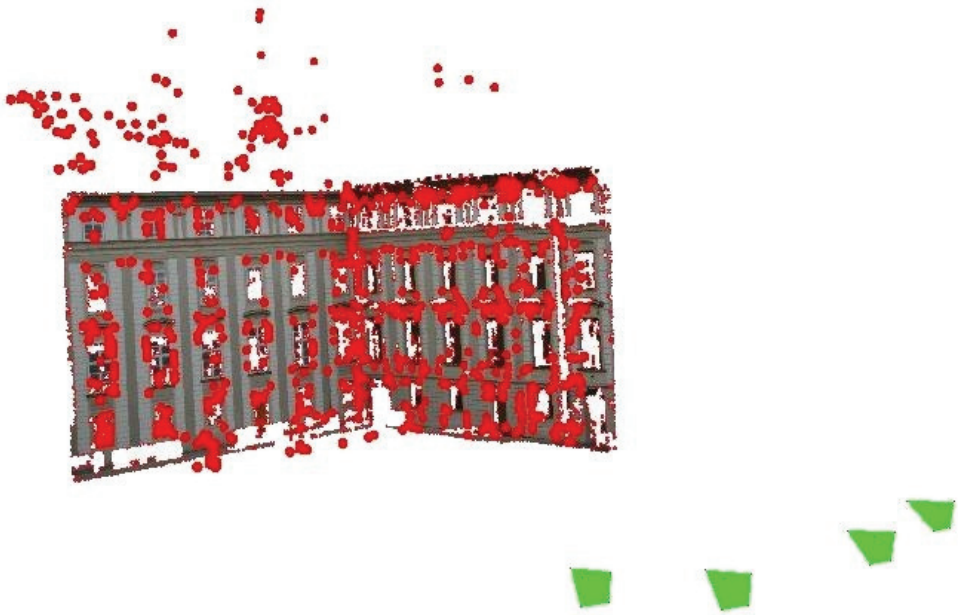

Figure 2: 3D points (red), cameras (green pyramids) and façade planes including holes of areas not on the façade planes generated from four images of Prague's Hradschin

3. APPEARANCE BASED OBJECT EXTRACTION

The ultimate goal of image understanding is to fully automatically determine, what can be seen where. Basically, this rests on huge amounts of knowledge, with Popper (1999) arguing, that most of our knowledge being 'inborn' knowledge. This has become clear also empirically by the difficulty or even impossibility to comprehensively model the real 3D world manually from scratch.
Appearance based object extraction is a way to tackle two aspects of the above problems: Parts of the model are learned and the complexity of the 3D model is avoided by solving the problem in

image space. The basic idea of recent approaches on appearance based object extraction is to describe an object by image patches around salient points and their relations in the image. Which patches and which relations characterize an object are learned from training data.
Particularly, Agarwal et al. (2004) employ (normalized) cross correlation (CC) to compare image patches around Förstner points (Förstner and Gülch 1987) to find cars seen from the side. The image patches of the training images are clustered (e.g., many wheels look similar) and for the clusters the relative locations in terms of direction and distance to other patches are learned. The recognition of a car is mapped to the problem of deciding if a part of an image contains a car. For a particular image part, Förstner points are extracted, the patches around the points are compared to the clusters via CC, for the similar patches the relations are computed and, based on them, it is decided how likely it is, that a car is present or not. To locate a car in an image, it is split into parts with approximately the size of the car, for all parts the likelihood is determined and finally a car is deemed to be detected at all positions with a likelihood higher than its surroundings and above a given threshold. To be able to deal with images of different resolutions, an image pyramid is used and a hypothesis for a car has to be a maximum in the spatial as well as the scale domain.
An approach based on an implicit shape model with better results and which is also conceptually superior than the one above was proposed by Leibe et al. (2004). Again, CC is used to build clusters, but scale-invariant features are used (Leibe and Schiele 2004) and the hypotheses for objects are found via generalized Hough transform (Ballard 1981). Additionally, the object is segmented by back projecting the learnt image patches into the image. The approach obtains for the task to find a large number of cars an equal error rate of more than 97% for the fixed-scale and 91% for the scale-invariant solution.
The simple core of the implicit shape model with the points, the CC, and the generalized Hough transform has led us to the idea, to use it for façade interpretation, particularly, window detection. Here, we have the big advantage, that the viewpoint is relatively limited as we can project the façades on their corresponding planes and rotate them to the vertical direction. Additionally, it is helpful to normalize the pixel size on the façades. Figure 3 shows eleven out of 72 image parts used for learning. The red points were detected with the Förstner operator with a fixed set of parameters used also for the actual object detection. The yellow points in the centers of the windows have been marked manually.

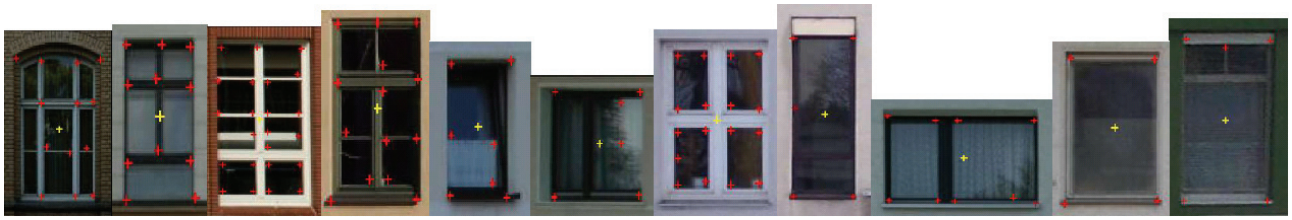

Figure 3: Eleven of 72 image parts used for learning with Förstner points (red) and manually determined centers of windows (yellow)

To detect windows, Förstner points are extracted (cf. Figure 4, left). Then, the patches around these points are compared by means of CC with all image patches learnt in training. If CC is above an empirically determined threshold of 0.8, the known difference vector for the learnt patch to the center point marked in yellow in Figure 3 is employed to increment hypotheses for window centers in an initially empty accumulator array (cf. Figure 4, right). Figure 4, right, shows, that there is a larger number of these hypotheses inside the windows than outside. This is due to the facts, that only patches which look similar to training patches trigger hypotheses and, that, while some patches, as, e.g., the upper right corners of an upper and a lower part of a window, might give ambi-

guous evidence, usually only evidence for correct hypotheses clusters. To actually end up with one hypothesis per window, we employ the fact, that windows have a certain size. Therefore, the accumulator array is integrated, i.e., blurred with a Gaussian with an appropriate σ, and the maxima beyond an empirically determined threshold are taken as hypotheses for centers of windows (cf. Figure 5). Please note, that none of the windows in Figure 4 has been used for training.
While the solution for our problem looks very promising (cf. also further results given in the next Section), there are limitations for this approach for more general problems such as building detection. Particularly, the 3D geometry is modeled only implicitly and the shape is restricted to the learnt examples. For this an idea might be to learn regular parts of the more complex generic objects. To improve the modeling of the image function also in terms of its invariance concerning viewing angle and scale, Lowe's (2004) SIFT operator should be a high quality alternative to CC.

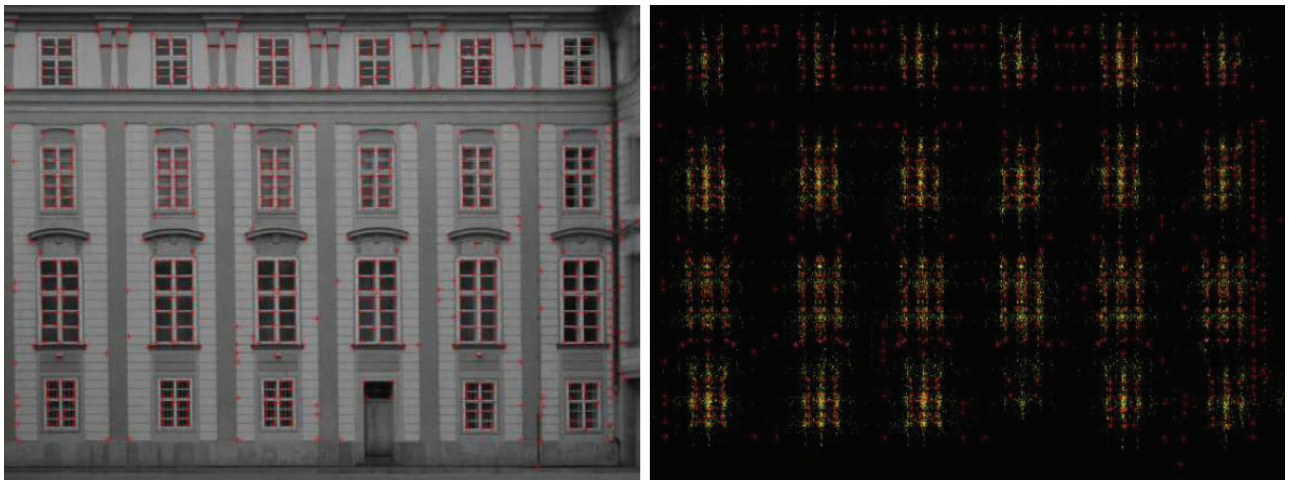

Figure 4: Façade (left) and accumulated evidence for window centers (right), both with Förstner points (red)

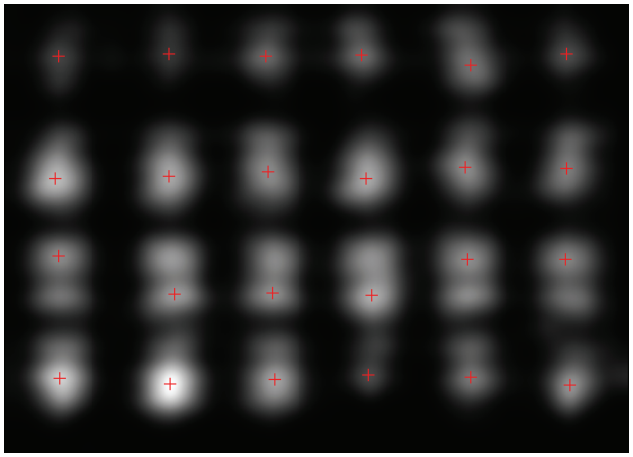

Figure 5: Accumulated evidence for window centers integrated with a Gaussian filter and their maxima (red crosses)

4. STATISTICAL AND COMBINATORIAL MODELING

We made our first experiences with non-deterministic modeling when using RANSAC (cf. Section 2). Defining the output of an algorithm also in terms of the probability for a correct solution feels unsafe in the beginning. Yet, if one can solve problems efficiently, where the other option is to wait unacceptably long, one starts to accept this change of attitude and considers it as a means for different kinds of problems.

A very recent and advanced example is (Tu et al. 2005), where segmentation of image regions is linked with the recognition of semantic objects such as faces or text. The basic idea is to determine probabilities for different segmentations and objects and their maxima by means of statistical sampling employing MCMC. A very important development for this, which has made MCMC feasible for a much larger class of problems, because the number of parameters of the problem is allowed to change dynamically, is the extension by means of reversible jumps (RJ) to RJMCMC (Green 1995).

The latter is employed in work closer to the realm of photogrammetry by Stoica et al. (2004) and particularly Dick et al. (2004). Both show an extremely important feature of MCMC based modeling, namely the simulation of the given knowledge by sampling into the prior distribution. E.g., in (Stoica et al. 2004) it is shown how realistically looking road networks can be simulated by starting from one given segment and priors for lengths and spatial relations such as relative angles. Similarly, Dick et al. (2004) generate a variety of buildings from given simple buildings and distributions for window, door, etc., sizes and the location of windows, etc., on façades. The ability to simulate is a big advantage compared to modeling tools such as semantic nets, where the correctness of the model can only be guessed from results of the extraction from given data.

To actually extract objects, the prior knowledge is linked to the likelihood, that an object is actually present in the data. Often the latter is generated by comparing a visualization of the hypothesized model with the original data. The posterior distribution is obtained as the usual multiplicative combination of prior and likelihood.

To extract windows on façades, we start with the hypotheses for window centers generated by the appearance based approach presented above and hypothesize windows in the form of black squares with a relatively small size on a gray background. For a more correct modeling of the actual image including disturbances, noise with ± 10 gray values is added. As real windows consist of finer scale details, abstraction is used in the form of gray-scale opening-closing scale-space filtering realized by the Dual Rank filter described in (Eckstein and Munkelt, 1994). The original ortho-projected façade image, the abstracted ortho image, as well as the simulated image are given in Figure 6.

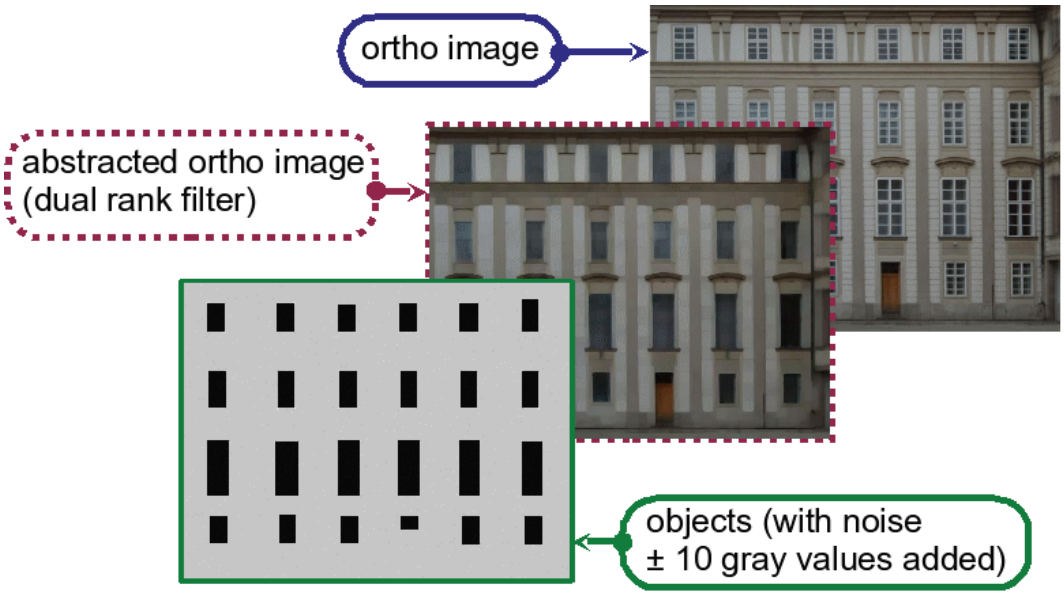

Figure 6: Abstraction hierarchy consisting of the original ortho image, the ortho image abstracted via Dual-Rank filter, as well as the final MCMC-model with added noise

For the comparison of the simulated and the abstracted ortho image again the CC is used. Here the result is interpreted as the likelihood of a hypotheses for being a window. In the simplest case we

just use priors for the width and the height of windows and change the width, height, as well as the location, i.e., the x- and the y-position randomly (Monte Carlo). That the values for an iteration depend only on the directly preceding iteration is the Markov property of the Markov Chain. As we know from the above appearance based approach more about the location than about the width and the height, we change the latter more often. The final goal is the solution with the highest posterior.
To avoid to get stuck in local maxima, we use simulated annealing. Here this means, that according to a (temperature) parameter, which is high at the beginning, also solutions are accepted which are worse than the solutions obtained before. With each iteration the temperature and, therefore, the probability to accept a solution worse than the one obtained before is reduced leading finally ideally to the global, but in realistic cases to a suitable maximum. The result for this is shown for the running example in Figure 7. All windows, marked in green have been found (hypotheses given as white squares) and have been delineated rather well. Further results are given in Figure 8. Again, all windows have been found, though one can see some deficits of the outline especially closer to the margin of the image.

Figure 7: Results of MCMC – Hypotheses (white squares) and windows (green rectangles)

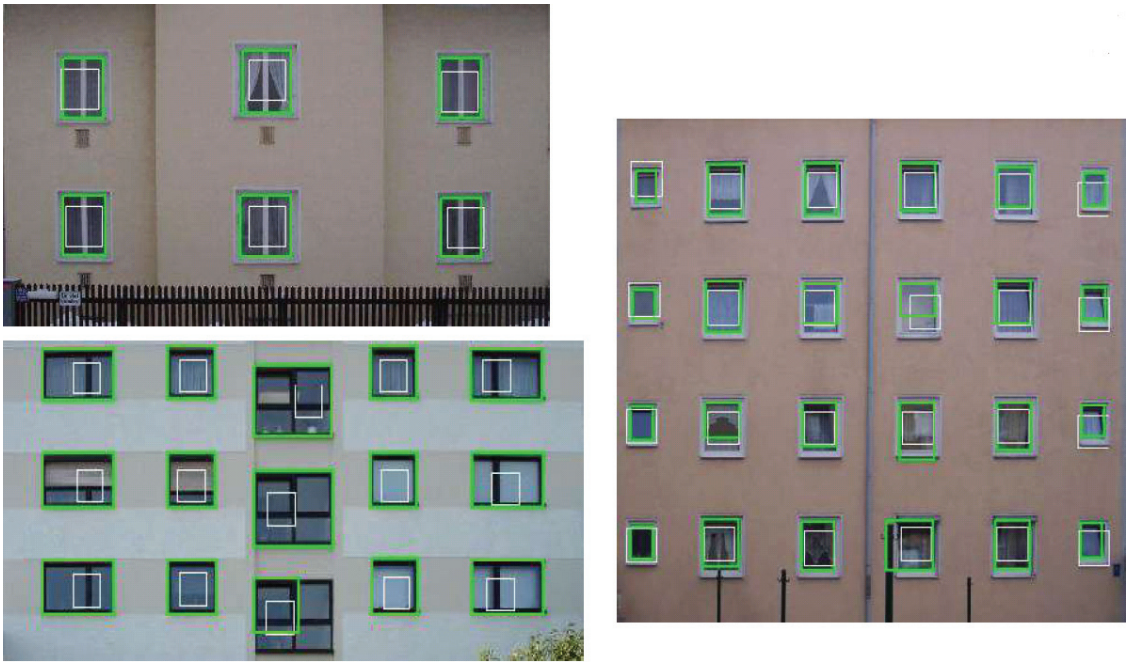

Figure 8: Additional results – Hypotheses (white squares) and windows (green rectangles) for buildings from Bonn and Munich

While MCMC only allows to model objects with a given fixed number of parameters, by means of RJMCMC it becomes possible to switch between different types of models with different numbers of parameters. Here, these are, e.g., modeling individual windows, or rows, columns, or even grids of windows. In Figure 9 preliminary results for the determination of rows of windows via RJMCMC are shown. Though the regularity induced by the row helps in the determination of the location of windows, and potentially makes it possible to detect also partially hidden windows, it can also lead to new problems such as interpreting the window above the door in the lower center of Figure 9 together with parts of the door as evidence for a regular window. Here, the evaluation of the likelihood has to be improved and also the prior probability of doors on the ground floor should be helpful.

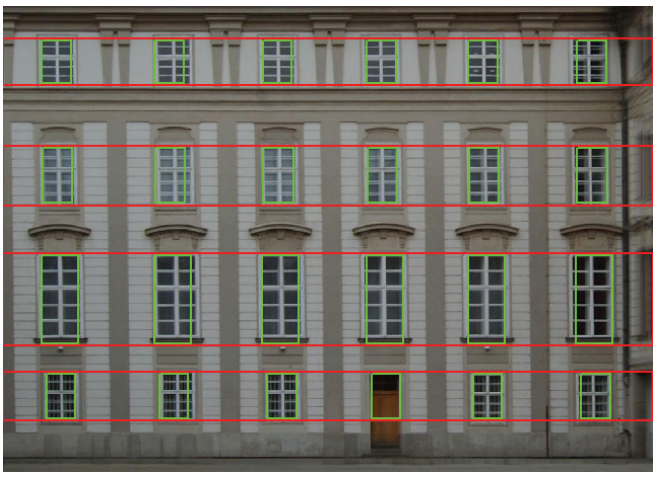

Figure 9: Rows of windows generated by means of RJMCMC

5. CONCLUSIONS

We have tried to take up the challenge issued by the organizers of the photogrammetric week by focusing on two topics of image understanding with broad recent interest. These are particularly appearance based and statistical approaches for which a large potential is seen to model objects as one can soundly combine prior knowledge with image information. Additionally, recent progress in automatic 3D reconstruction from uncalibrated imagery has been demonstrated.
In spite of the large scientific progress we have seen over the last fifteen years, still every new solution opens up more than one question. This is scientifically extremely interesting, but widens the gap between what people would like to have and what can actually be achieved.

6. REFERENCES

Agarwal, S., A. Awan and D. Roth (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.

Ballard, D., (1981): Generalizing the Hough Transform to Detect Arbitrary Shapes, Pattern Recognition, 13(2):111–122.

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

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

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

Fischler, M. and R. Bolles (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 E. Gülch (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.

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

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

Leibe, B., A. Leonardis and B. Schiele (2004): Combined Object Categorization and Segmentation with an Implicit Shape Model, ECCV 04 Workshop on Statistical Learning in Computer Vision, pp. 1–15.

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

Lowe, D.G. (2004): Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(1), 91–110.

Mayer, H. and S. Reznik (2005): Building Façade Interpretation from Image Sequences. Accepted for Joint Workshop of ISPRS and the German Association for Pattern Recognition (DAGM) on "Object Extraction for 3D City Models, Road Databases and Traffic Monitoring - Concepts, Algorithms, and Evaluation", Vienna, Austria.

Mumford, D. (2000): The Dawning of the Age of Stochasticity, Mathematics: Frontiers and Perspectives – American Mathematical Society, pp. 1–23.

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

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

Popper, K.R. (1999): All Life is Problem Solving. Routledge, London, Great Britain.

Stoica, R., X. Descombes and J. Zerubia (2004): A Gibbs Point Process for Road Extraction from Remotely Sensed Images, International Journal of Computer Vision 57(2): 121–136.

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