Photogrammetric Week '15 Dieter Fritsch (Ed.) Wichmann/VDE Verlag, Belin & Offenbach, 2015

From Orientation to Functional Modeling for Terrestrial and UAV Images

Helmut Mayer, Neubiberg

ABSTRACT

While images have been acquired from planes, satellites or the ground since decades, only recently the acquisition from Unmanned Aerial Vehicles (UAVs) has become an additional option. This paper focuses on the automatic processing of images from small UAVs together with terrestrial images from orientation to the functional modeling of buildings. It is shown how highly precise orientation can be achieved fully automatically only based on images also for large unordered sets of wide baseline images. The orientation is used for high quality fully 3D reconstruction from possibly many high resolution images employing efficient statistical modeling in voxel space based on octrees and a Total Variation (TV) based feature for the quality of the disparities. From dense 3D reconstruction functional models are derived in the form of façades, roofs, windows and doors using statistical sampling. Results for several datasets give an impression of the potential of the developed processing chain.

1. INTRODUCTION

Research towards the automatic orientation of unordered sets of images with general orientations dates back to the mid of the nineties of the twentieth century. Particularly, Pollefeys et al. (2002, 2008) have made major contributions concerning matching as well as (relative) orientation and dense 3D reconstruction. Arguably a major breakthrough has been (Frahm et al., 2010) as a successor of (Agarwal et al., 2009) who have shown that huge image sets of millions of images from ‘Community Photo Collections’ (Goesele et al., 2010) can be analyzed concerning overlap and tens of thousands of images oriented on one standard PC on one day.

Key to these developments have been the 5-point algorithm of Nistér (2004) and RANdom SAmple Consensus – RANSAC by Fischler & Bolles (1981). The former allows for a direct solution of the relative orientation problem of a (possibly weakly) calibrated image pair. Here, ‘direct’ means, that the image can be oriented even though no approximate values are available. RANSAC allows to deal with wrong matches, particularly common for wide baselines. Because it takes into account knowledge about the solution, namely that the precision of correct matches is approximately known in terms of the error in pixels, it can deal with much less than 50% correct observations, the theoretical limit for a robust approach not considering this knowledge.

In Section 2. of this paper we present an approach based on (Bartelsen et al., 2012, Mayer et al., 2012, Mayer, 2014) which focuses on large unordered sets of images taken from Unmanned Aerial Vehicles (UAVs) as well as from the ground with possibly large base lines. It combines 5-point algorithm and RANSAC from computer vision with highly precise least squares matching and robust bundle adjustment from photogrammetry. Additionally, we introduce very recent work on graph based derivation of minimum sets of triplets which allow for the orientation of large image sets with wide baselines in a comparably short amount of time in comparison to state-of-the-art algorithms such as VisualSFM (Wu, 2011, 2013).

Dense reconstruction from multiple images has received much attention in photogrammetry, e.g., (Ebner et al., 1987). Yet, only recent work such as Semi Global Matching (SGM) by Hirschmüller (2008) have demonstrated, that based on high quality digital imagery as well as the high overlap possible with recent digital aerial cameras, reliable and precise pixel wise disparity computation is possible. We have demonstrated in (Kuhn et al., 2013, 2014) that SGM can be extended to fully 3D reconstruction even for large sets of high resolution images employing local optimization based on statistically coherent volume modeling, efficient octree data structures and a Total Variation (TV) based feature for the quality of the disparities. This work is summarized in Section 3.

Dense 3D models are the basis for functional modeling, in this case the generation of models for buildings consisting of their façades, roofs, windows and doors (cf. Section 4.). The classical work of (Dick et al., 2004) was based on generative modeling from terrestrial images. Our earlier work (Mayer & Reznik, 2007, Reznik & Mayer, 2008) employed an Implicit Shape Model (ISM) (Leibe & Schiele, 2004). Objects, i.e., windows, were modeled in the form of small image patches and their relations. In our recent work (Nguatem et al. 2012, 2013, 2014) we use statistical sampling and model selection based on the dense 3D models from Section 3.

2. ORIENTATION FOR WIDE BASELINE IMAGE SETS

This section describes our approach for the orientation of large unordered sets of images taken from UAVs and from the ground, possibly with a wide baseline. We first (Section 2.1) describe the reconstruction of the relative geometry for pairs and triplets. The latter are useful, because the results for triplets are much more reliable. Triplets are then efficiently merged hierarchically (Section 2.2) to larger and larger image blocks. The first two sections describe an approach (Bartelsen et al., 2012, Mayer et al., 2012, Mayer, 2014) which can orient images when it is known, which images overlap. Yet, no knowledge about the relative geometry is needed. We extend this to the situation where the overlap is not known in Section 2.3. The representation for it is based on a very efficient approximate matching for all possible pairs and the reduction of the number of pairs and triplets to be oriented by the reliable but slow methods of Section 2.1 based on different graph structures. Results show the efficiency of our approach also in relation to recent state-of-the-art approaches such as VisualSFM (Wu 2011, 2013).

2.1. Orientation of pairs and triplets

A basic step for automatic orientation is the determination of conjugate points. The gold standard for point detection is the Scale Invariant Feature Transform – SIFT (Lowe, 2004). For pair and triplet orientation we only use the SIFT points, but not the feature description. The points including scale are detected only once per image. For every point there is a unique ID consisting of the image number and the point number in the image. This is particularly important for hierarchical merging (cf. next section). For hierarchical processing in terms of resolution, we simply threshold the points concerning scale.

The generation of conjugate points consists of normalized cross correlation followed by least squares matching (Förstner, 1982, Grün, 1985). The former is used to reduce the number of hypothesis for the time consuming least squares matching, the latter makes available covariance information employed in the latter stages to obtain a more reliable and precise result.

Conjugate points are the input to relative orientation based on RANSAC, the 5-point algorithm and Expectation Maximization (EM). The Geometric Robust Information Criterion – GRIC (Torr, 1997) is employed instead of the standard counting of inliers in RANSAC. Via GRIC it is possible to take into account if an inlier is closer or further away from the correct solution, i.e., the epipolar line, as well as how precise the match is. For the latter the covariance information from least squares matching is employed.

EM is used to deal with a similar situation as locally optimized RANSAC (Chum et al., 2003), namely, that a solution consisting only of inliers is not necessarily precise enough to comply with most or even many of the inliers. Thus, one iterates between Expectation, i.e., the classification of inliers, and Maximization, i.e., improvement of the parameters, in this case by means of robust bundle adjustment.

Once the relative orientation of pairs has been determined for relatively few points on a low resolution level of the pyramid in the range of 150 pixels, the epipolar lines are known. Thus, matching triplets of points in higher resolution, i.e., with more points becomes computationally feasible. Also for triplets, GRIC, RANSAC and EM are employed. Geometrically, we employ the 5-point algorithm two times for images 1 and 2 as well as 1 and 3 and then determine the relative scale in the form of the median of the distance ratios to the five points in the two pairs.

2.2. Hierarchical merging

Serial merging of triplets into an image block (Bartelsen et al., 2012, Mayer et al., 2012) becomes extremely time consuming for larger blocks, because it was found to be necessary to bundle adjust the results after every addition of a triplet. Thus, in (Mayer, 2014) a much more efficient hierarchical merging has been proposed. The basic idea is to employ the unique ID of points (cf. the previous section). By this means finding conjugate points in triplets or image (sub) blocks is reduced to finding the same ID which can be done extremely efficiently and avoids the deletion of redundant 3D points in the earlier approach.

The merging is done hierarchically. Instead of adding just one image, image blocks are merged with an overlap of two images. The latter makes it possible to determine the 3D similarity transformation based only on the images with no need to rely on possibly unreliable 3D points. In the ideal case, two blocks with n images lead to a new block of n * 2 – 2 images. E.g., starting with 3, one obtains blocks with 4, 6, 10, 18, 34, 66, 130, 258, 514, 1026, etc. images. Bundle adjustment is just employed for the generated image blocks and besides the last block, all other blocks are independent and the merging and bundle adjustment can, thus, be computed in parallel.

An additional gain in efficiency is achieved by reducing the number of points per image. This number has to be high for pairs and triplets to guarantee stable image blocks also for small overlaps. Yet, once the overlap is known, much less points are needed. While we had had the intuition, that points seen in many images as well as a good distribution of points should be especially beneficial, experiments based on statistical tests using the Mahalanobis distance proved that only a totally random deletion of (3D) points leads to unbiased results (Mayer, 2014).

2.3. Automatic determination of image block

The preceding sections have shown means to orient images if nothing about the images is known but the camera calibration as well as which images overlap (but not how). To become even more flexible, i.e., to be able to deal with unordered image sets with only an approximate camera calibration (e.g., from the Exchangeable Image File Format – Exif tags of the images), we have developed an extended approach. It determines the overlap extremely efficiently and produces a limited set of hypotheses for overlapping pairs and triplets to be validated by our reliable but slow methods for pairs and triplets based on 5-point algorithm, RANSAC, GRIC and EM introduced in Section 2.1 above.

While normalized cross correlation is a good means to find matches also for wide baselines, their computational efficiency is low. We, therefore, use a variant of SIFT optimized for highly parallel Graphical Processing Units – GPUs named SiftGPU (Wu, 2007), to generate SIFT points and feature vectors consisting of 128 real values for matching. Yet, instead of computing the (continuous) distance between the feature vectors, we embed the feature vectors and obtain 128 bit values which can be extremely efficiently compared based on the Hamming distance, i.e., common zeroes and ones.

SIFT matching gives us the number of matches between pairs of images. This is the basis for the determination of an approximately minimal set of images pairs and triplets which link as many images in the unordered set as possible. For the pairs this is based on Minimum Spanning Trees

(MSTs). As verification of pairs based on the methods introduced in Section 2.1 can show that a hypothesized link is not valid, not one but several MSTs are obtained. For triplets a similar concept to MSTs are terminal Steiner minimal trees (Lin & Xue 2002). They are employed to obtain an approximately minimal set of triplets linking all images which are then verified by the method for triplet orientation of Section 2.1. From the oriented triplets one or several image blocks are built based on hierarchical merging (cf. the previous section).

First results show that our approach is fast also in comparison with state-of-the-art approaches such as VisualSFM (Wu, 2011, 2013). For the data set presented in Figure 1 consisting of 655 images from the ground and from several UAVs our approach took just 32 minutes, while SFM needed 271 minutes on the same hardware. The data set in Figure 2 consists of 1570 images. Our approach took 186 minutes and VisualSFM 1674 minutes. The number of derived points are similar, but due to the use of least squares matching, we obtain a much more precise result: 0.42 pixels instead of 2.25 pixels for VisualSFM for the first data set and 0.29 pixels instead of 2.04 pixels for the second.

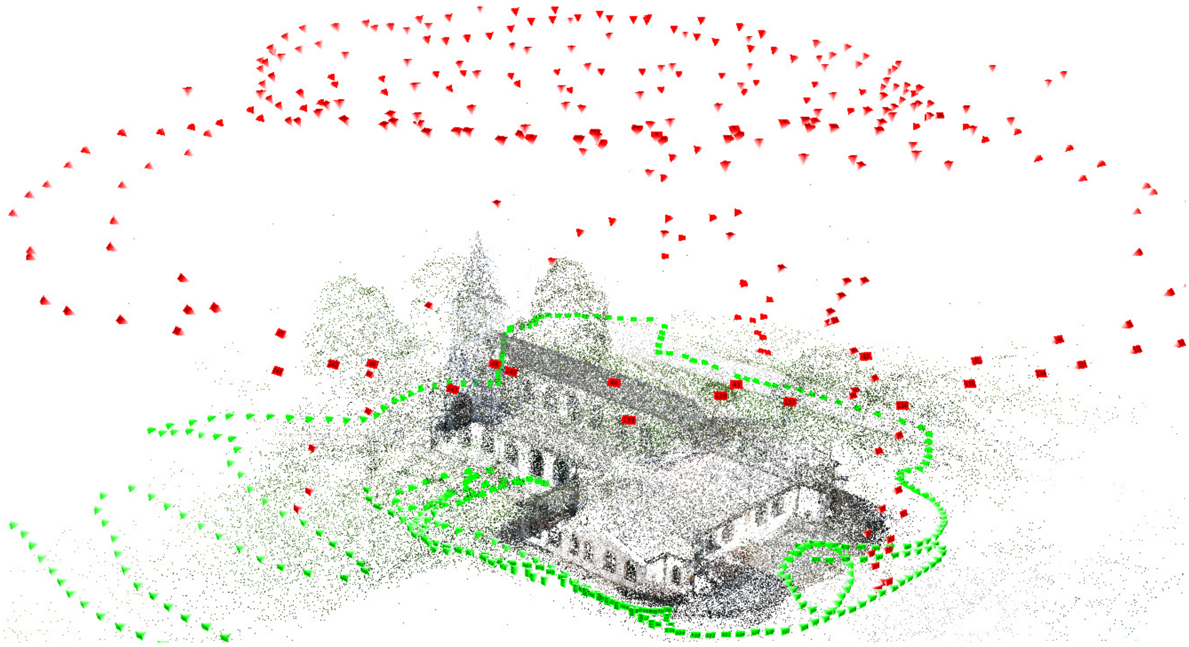

Figure 1: Image set Church (of Bundeswehr University Munich) consisting of 655 images taken by multiple cameras from the ground and a UAV visualized by colored pyramids.

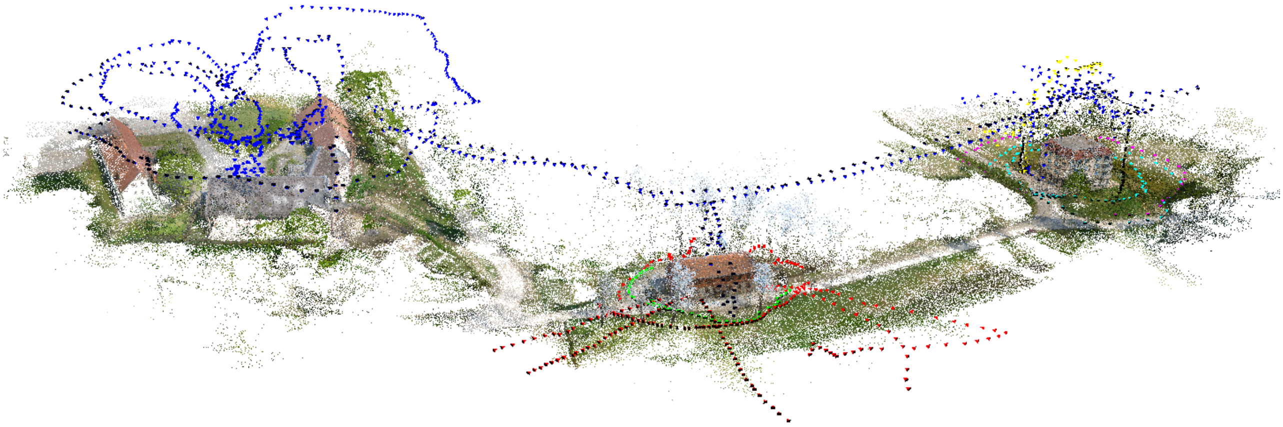

Figure 2: Image set Village consisting of 1570 images taken by multiple cameras from the ground and several UAVs.

3. FULLY 3D DENSE RECONSTRUCTION

Our work on fully 3D reconstruction (Kuhn et al., 2013, 2014) is based on Semi Global Matching (SGM) by Hirschmüller (2008). While global approaches such as (Vu et al., 2012) produce the results with the highest quality, their scalability towards a large number of high resolution images is limited. Our local volumetric approach is based on the seminal work on range image integration by Curless & Levoy (1996), but reinterprets and extends it by a novel probabilistic formulation. We, particularly, have analyzed the 3D error taking into account a sound stereo error model and have developed an approach for the fusion of data with strongly varying quality.

The stereo error model is usually dominated by the uncertainty of the disparity which is not known and can vary from one tenth to a couple of pixels. We have developed a Total Variation (TV) based feature describing how large the area is in which the image function varies less than a given threshold. This feature was found to be highly correlated with the quality of the disparity. The corresponding function is learned by an EM approach based on reference data.

To achieve nearly unlimited scalability, the volume data is represented in an octree which is dynamically generated and split into subtrees each handled on a single computing node. The data in the subtrees can be processed in parallel on as many nodes as available. Because the volumes of the subtrees overlap by the size of the influence area of the employed local probabilistic approach, the results for the subtrees can be seamlessly integrated without need for further editing.

Results show that our approach generates high quality results also for data sets with strongly varying intersection geometry and resolution on the object. While the overall complexity is high, the computing times are limited when distributing the computation on a large cluster with more than 100 computing nodes. This demonstrates a good scalability. Figure 3 presents results for the Church of Bundeswehr University Munich, Germany. Particularly for the door on the right side small details such as the door handle are represented in 3D. Overall, the results are smooth and yet detailed with few holes besides for areas occluded in all or all but one image.

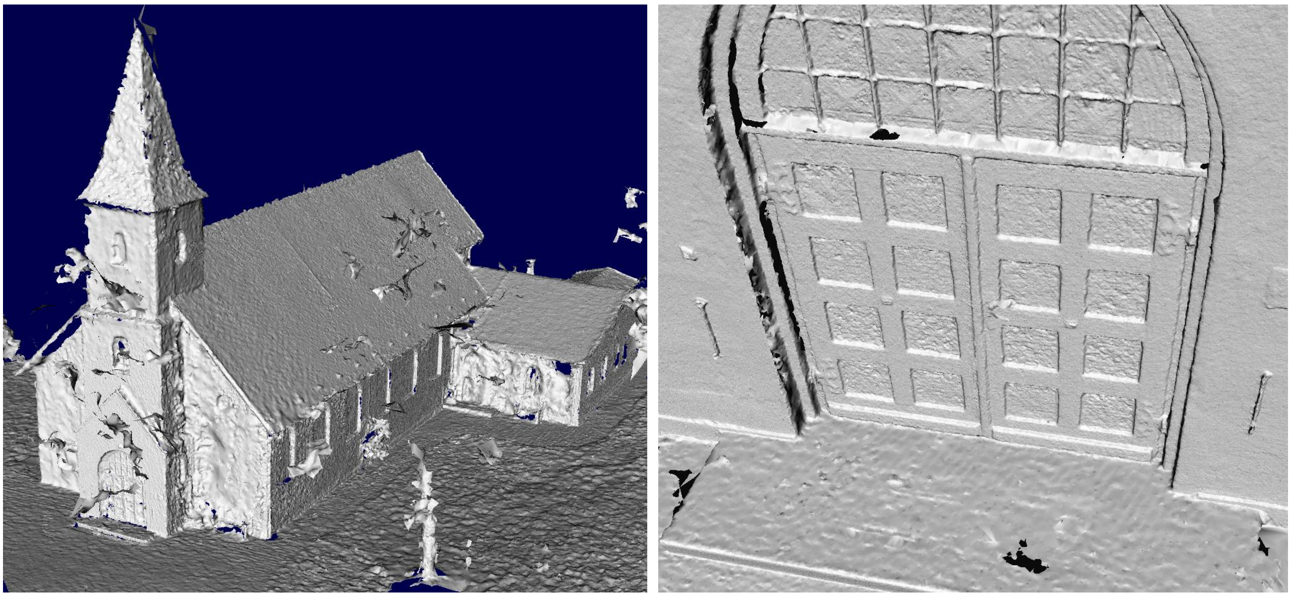

Figure 3: Dense 3D reconstruction of the Church of Bundeswehr University Munich (cf. orientation in Figure 1) – full model (left) and detail for door (right).

4. BUILDING RECOGNITION AND FAÇADE INTERPRETATION

For functional modeling we start with the determination of the façades of buildings in the form of vertical planes (Nguatem et al., 2012). They are the basis for the reconstruction of the roof (Nguatem et al., 2013) and the localization of windows and doors (Nguatem 2014).

To estimate the vertical direction (Nguatem et al., 2012), we make use of the fact that in our data which consists of images taken from the ground and UAVs one can see the roofs, but also the façades. Usually the area of the latter is much larger and, thus, we compute a first approximation of the vertical direction based on the cross products of the normals of pairs of points on the dense 3D surface (for which the majority lies on the façades). The vertical direction is refined based on edges in the point cloud and employed to estimate vertical planes with a RANSAC based approach. Because the vertical direction is already known, two randomly sampled points are sufficient to determine a vertical plane. For the resulting (infinite) planes the outlines are estimated with line sweeping. Finally, adjacent façades are determined based on the vicinity of their outlines and intersected to obtain a closed outline for the building.

The façades are the basis for the reconstruction of the roof using statistical sampling (Nguatem et al., 2013). Because the gutter height can only be determined approximately from the façade outlines, it is estimated together with the other parameters of the roof. Of particular importance is the comparison between different roof types, such as flat, gable, hipped or mansard, described by parametric models. It is solved via model selection based on the weights for the maximum a posteriori (MAP) solutions obtained by statistical sampling in a Bayesian setting. For improved efficiency, a coarse to fine scheme is employed.

Figure 4 presents on the left side a dense 3D model of the building on the right side of Figure 2. The balconies, their supporting posts as well as the roof overhang show that the representation is fully 3D. On the right side of Figure 4 a 3D functional model derived from the dense model is shown. It consists of the façades, parallel planes for the balconies as well as smaller parts of the façades in front of the main façades, and a mansard roof determined by means of model selection.

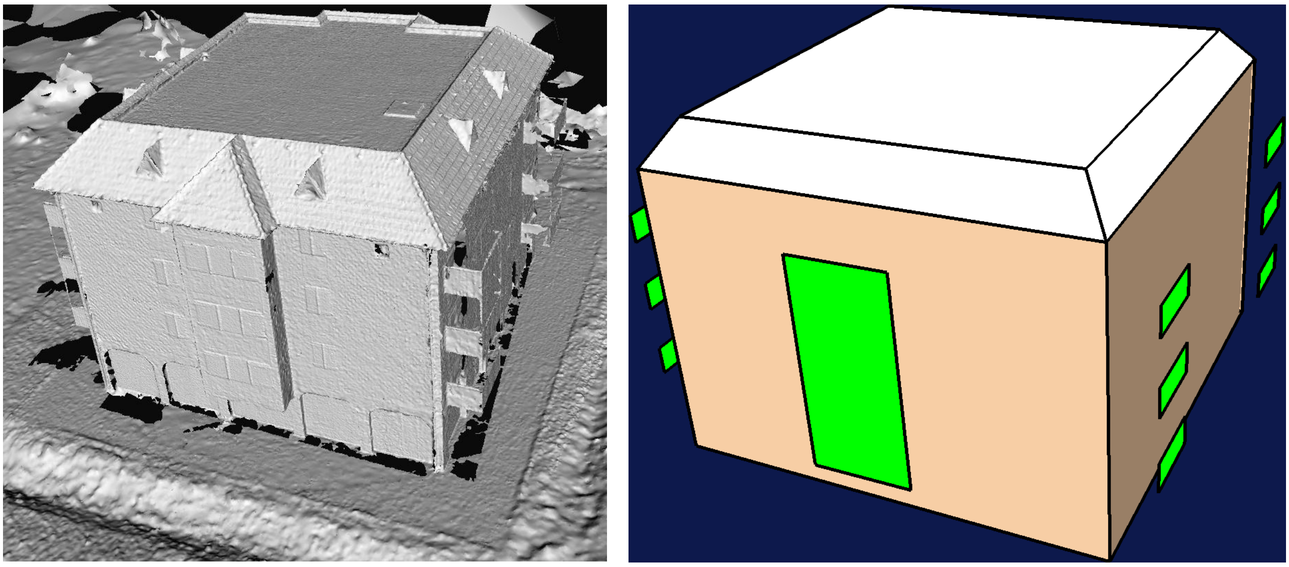

Figure 4: Dense 3D reconstruction of right building in Figure 2 (left) and 3D model derived from it (right).

For the façades doors and windows are determined (Nguatem et al., 2014) starting with the generation of several segmentations based on so-called ‘plane-angular sweeps’. I.e., the given hypotheses for the façade plane are shifted slightly along their normal directions and rotated away from the vertical direction. Windows and doors are modeled by means of B-splines modified by affine transformations. Like for the roofs, the distinction between different architectural types of doors and windows, such as rectangular, half-circular or gothic, is based on model selection for the MAP estimates obtained by probabilistic template matching between the models and the segmentations. By means of clustering the search space, the approach can be run in parallel for disjoint parts of the segmentation.

Figure 5 shows on the left the point cloud determined by fully 3D reconstruction (Section 3.). On the right, the detected façade, windows and doors are presented. All windows and doors besides the two windows on the very left have been found. The outlines are mostly correct. Figure 6 presents the dense point cloud as well as windows and doors for several façades. All windows besides the three on the upper right have been found. Their outlines are for the largest part correct. Not all doors have been detected and their outlines are only partly correct also due to occlusions by other objects.

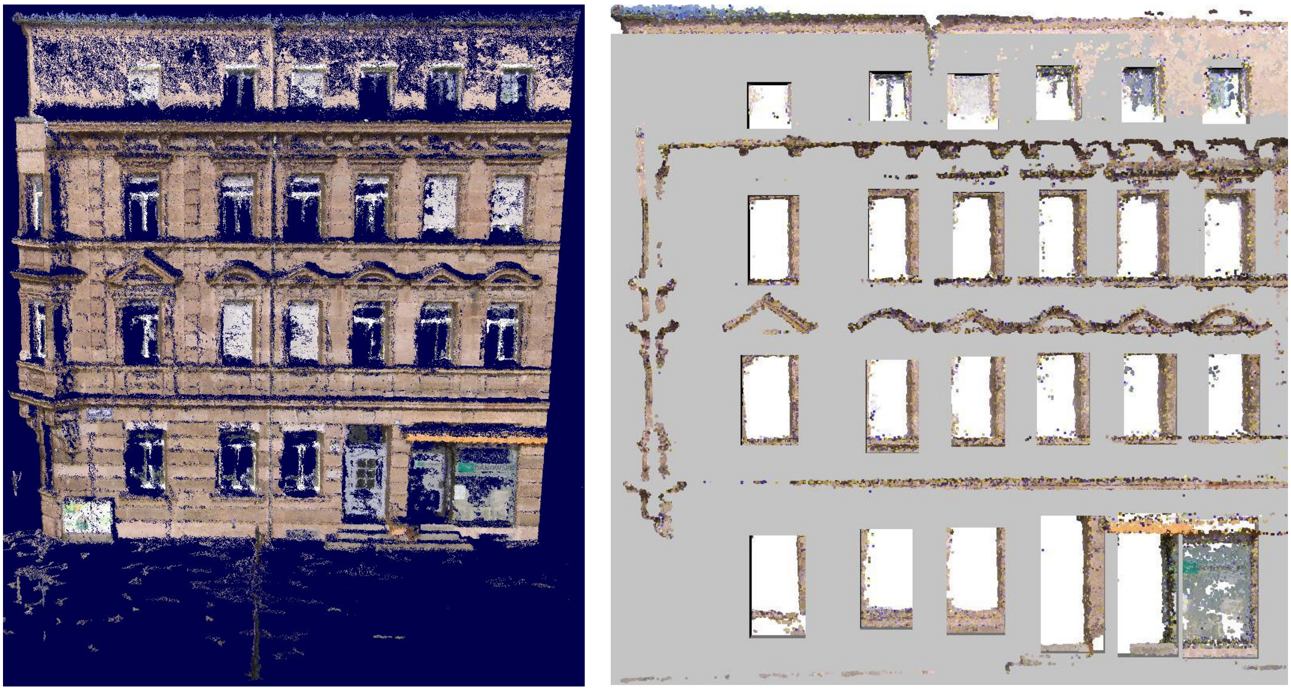

Figure 5: Dense 3D point cloud (left) and windows as well as doors as holes in the façade plane (right).

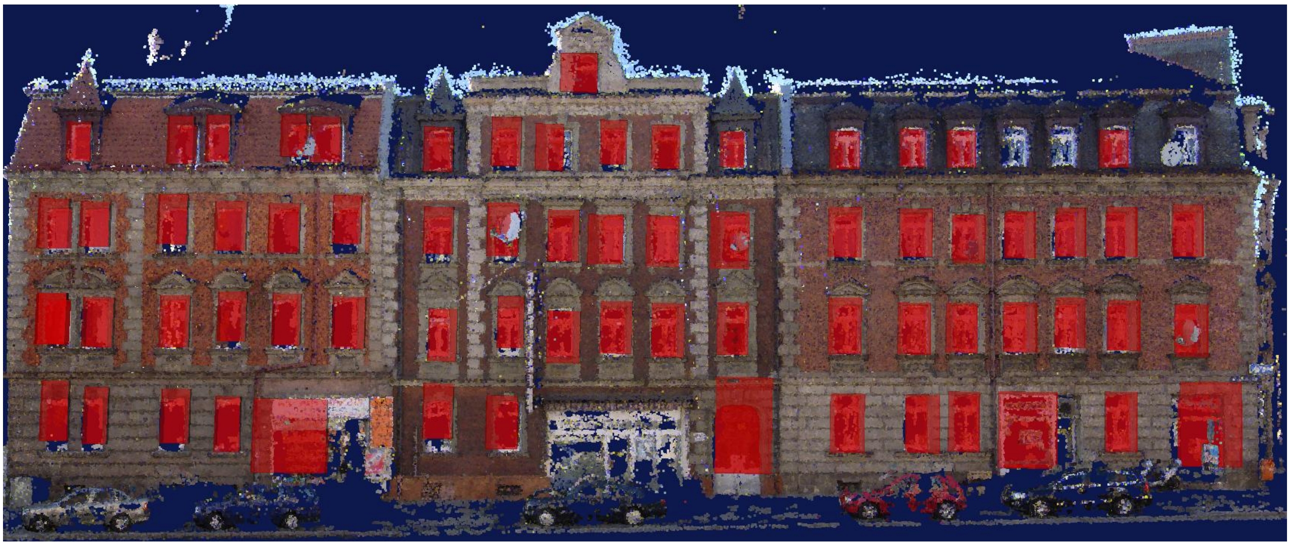

Figure 6: Dense point cloud of several façades with windows and doors marked by red boxes.

5. CONCLUSION

We have presented an approach for images taken from the ground and from UAVs ranging from camera orientation to the functional modeling of buildings generating models consisting of façades, roofs, windows and doors. For orientation we combine highly robust approaches from computer vision with the high precision of photogrammetry. Results show a strongly improved precision as well as speed in comparison with a state-of-the-art approach. Our approach for the generation of dense fully 3D models reaches a very high quality by means of a statistically sound modeling of the voxel space and is suitable for a large number of high resolution images, distinguishing it from most other approaches available. Finally, we have shown how to employ the generated dense fully 3D model for the determination of façades, roofs windows and doors based on a consistent Bayesian statistical modeling allowing for model selection between different roof types and architectural styles for windows and doors.

Concerning future work, for orientation the determination of weak pairs, e.g., by means of (Michelini & Mayer, 2014), could help to stabilize the generation of the image block. For dense fully 3D reconstruction, the introduction of the estimated full covariance information about the camera orientations could be a, though computationally intensive, means to improve the quality of the surfaces for geometrically difficult configurations. Finally, functional modeling of façades could be extended by ornamental objects such as cornices (Brandenburger et al., 2013) and the organization of the façade in terms of a grid, rows or columns of windows could be taken into account.

6. ACKNOWLEDGEMENTS

I sincerely thank Martin Drauschke, Heiko Hirschmüller, Andreas Kuhn, Mario Michelini, and William Nguatem for their contributions.

REFERENCES

Agarwal, S., Snavely, N., Simon, I., Seitz, S. & Szeliski, R. (2009): Building Rome in a Day. International Conference on Computer Vision, pp. 72-79.

Bartelsen, J., Mayer, H., Hirschmüller, H., Kuhn, A. & Michelini, M. (2012): Orientation and Dense Reconstruction from Unordered Wide Baseline Image Sets. Photogrammetrie – Fernerkundung – Geoinformation (PFG), 2012 (4), pp. 421-432.

Brandenburger, W., Drauschke, M. & Mayer, H. (2013): Cornice Detection Using Façade Image and Point Cloud. Photogrammetrie – Fernerkundung – Geoinformation (PFG), 2013 (5), pp. 511-521.

Chum, O., Matas, J. & Kittler, J. (2003): Locally Optimized RANSAC. Pattern Recognition – DAGM, pp. 249-256.

Curless, B. & Levoy, M. (1996): A Volumetric Method for Building Complex Models from Range Images. SIGGRAPH '96 – 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 303-312.

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

Ebner, H., Fritsch, D., Gillessen, W. & Heipke, C. (1987): Integration von Bildzuordnung und Objektrekonstruktion innerhalb der digitalen Photogrammetrie. Bildmessung und Luftbildwesen, 5/87, pp. 194-203.

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), pp. 381-395.

Förstner, W. (1982): On the Geometric Precision of Digital Correlation. International Archives of Photogrammetry and Remote Sensing, (24) 3, pp. 176-189.

Frahm, J.-M., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S. & Pollefeys, M. (2010): Building Rome on a Cloudless Day. Eleventh European Conference on Computer Vision, IV, pp. 368-381.

Goesele, M., Ackermann, J., Fuhrmann, S., Klowsky, R., Langguth, F., Muecke, P. & Ritz, M. (2010): Scene Reconstruction from Community Photo Collections. IEEE Computer, 43 (6), pp. 48-53.

Grün, A. (1985): Adaptive Least Squares Correlation: A Powerful Image Matching Technique. South African Journal of Photogrammetry, Remote Sensing and Cartography, 14 (3), pp. 175-187.

Hirschmüller, H. (2008): Stereo Processing by Semi-Global Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30 (2), pp. 328-341.

Kuhn, A., Hirschmüller, H. & Mayer, H. (2013): Multi-Resolution Range Data Fusion for Multi-View Stereo Reconstruction. German Conference on Pattern Recognition – GCPR, pp. 41-50.

Kuhn, A., Mayer, H. Hirschmüller, H. & Scharstein, D. (2014): A TV Prior for High Quality Local Multi-View Stereo Reconstruction. 2nd International Conference on 3D Vision (3DV), pp. 65-72.

Leibe, B. & Schiele, B. (2004): Combined Object Categorization and Segmentation with an Implicit Shape Model, European Conference on Computer Vision, Workshop on Statistical Learning in Computer Vision, pp. 1-15.

Lin, G. & Xue, G. (2002): On the Terminal Steiner Tree Problem. Information Processing Letters, 84 (2), pp. 103-107.

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

Mayer, H. (2014): Efficient Hierarchical Triplet Merging for Camera Pose Estimation. German Conference on Pattern Recognition – GCPR, pp. 399-409.

Mayer, H., Bartelsen, J., Hirschmüller, H. & Kuhn, A. (2012): Dense 3D Reconstruction from Wide Baseline Image Sets. Outdoor and Large-Scale Real-World Scene Analysis – 15th International Workshop on Theoretical Foundations of Computer Vision, pp. 285-304.

Mayer, H. & Reznik, S. (2007): Building Façade Interpretation from Uncalibrated Wide-Baseline Image Sequences. ISPRS Journal of Photogrammetry and Remote Sensing, 61 (6), pp. 371-380.

Michelini, M. & Mayer, H. (2014): Detection of Critical Camera Configurations for Structure from Motion. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 40 (3/W1), pp. 73-78.

Nguatem, W., Drauschke, M. & Mayer, H. (2012): Finding Cuboid-based Building Models in Point Clouds, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 39 (B3B), pp. 149-154.

Nguatem, W., Drauschke, M. & Mayer, H. (2013): Roof Reconstruction from Point Clouds using Importance Sampling. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, II (3/W3), pp. 73-78.

Nguatem, W., Drauschke, M. & Mayer, H. (2014): Localization of Windows and Doors in 3D Point Clouds of Façades. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, II (3), pp. 87-94.

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., Nistér, D., Frahm, J. M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S. J., Merrell, P., Salmi, C., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewénius, H., Yang, R.,Welch, G. & Towles, H. (2008): Detailed Real-Time Urban 3D Reconstruction from Video. International Journal of Computer Vision, 78 (2-3), pp. 143-167.

Pollefeys, M., Verbiest, F. & Van Gool, L. (2002): Surviving Dominant Planes in Uncalibrated Structure and Motion Recovery. Seventh European Conference on Computer Vision, II, pp. 837-851.

Reznik, S. & Mayer, H. (2008): Implicit Shape Models, Self Diagnosis, and Model Selection for 3D Façade Interpretation. Photogrammetrie – Fernerkundung – Geoinformation (PFG), 3/08, pp. 187-196.

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

Vu, H. H., Labatut, P., Pons, J. P. & Keriven, R. (2012): High Accuracy and Visibility-consistent Dense Multiview Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34, pp. 889-901.

Wu, C. (2007): SiftGPU: A GPU Implementation of Scale Invariant Feature Transform (SIFT). cs.unc.edu/~ccwu/siftgpu.

Wu, C. (2011): VisualSFM: A Visual Structure from Motion System. ccwu.me/vsfm/.

Wu, C. (2013): Towards Linear-time Incremental Structure from Motion. 1st International Conference on 3D Vision (3DV), pp. 127-134.