ORIENTATION AND DENSE RECONSTRUCTION OF UNORDERED TERRESTRIAL AND AERIAL WIDE BASELINE IMAGE SETS

Jan Bartelsen[1], Helmut Mayer[1], Heiko Hirschmüller[2], Andreas Kuhn[12], Mario Michelini[1]

[1] Institute of Applied Computer Science, Bundeswehr University Munich, D-85577 Neubiberg, Germany.
(jan.bartelsen, helmut.mayer, andreas.kuhn, mario.michelini)@unibw.de
[2] Institute of Robotics and Mechatronics Oberpfaffenhofen, German Aerospace Center (DLR), D-82230 Wessling, Germany.
heiko.hirschmueller@dlr.de

KEY WORDS: Photogrammetry, Matching, Point Cloud, Aerial, Terrestrial, Urban, GPS/INS

ABSTRACT:

In this paper we present an approach for detailed and precise automatic dense 3D reconstruction using images from consumer cameras. The major difference between our approach and many others is that we focus on wide-baseline image sets. We have combined and improved several methods, particularly, least squares matching, RANSAC, scale-space maxima and bundle adjustment, for robust matching and parameter estimation. Point correspondences and the five-point algorithm lead to relative orientation. Due to our robust matching method it is possible to orient images under much more unfavorable conditions, for instance concerning illumination changes or scale differences, than for often used operators such as SIFT. For dense reconstruction, we use our orientation as input for Semiglobal Matching (SGM) resulting into dense depth images. The latter can be fused into a 2.5D model for eliminating the redundancy of the highly overlapping depth images. However, some applications require full 3D models. A solution to this problem is part of our current work, for which preliminary results are presented in this paper. With very small unmanned aerial systems (Micro UAS) it is possible to acquire images which have a perspective similar to terrestrial images and can thus be combined with them. Such a combination is useful for an almost complete 3D reconstruction of urban scenes. We have applied our approach to several hundred aerial and terrestrial images and have generated detailed 2.5D and 3D models of urban areas.

1 INTRODUCTION

Already (Pollefeys et al., 2000) have demonstrated, that sets of images from consumer cameras in combination with dense 3D reconstruction form a good basis for photo realistic visualization. (Pollefeys et al., 2002) presented one of the first approaches for relative orientation for a larger number of images in a general configuration, i.e., without known approximate values such as for aerial images. It employed uncalibrated images. This makes the approach very flexible, yet, on the other hand, reliant on sufficient 3D structure in the scene for the determination of calibration parameters.

While for the above work the overlap of the images is either known implicitly in the form of the order in a sequence, or explicitly, e.g., from an aerial flight plan. In (Schaffalitzky and Zisserman, 2002) one of the first methods which can automatically determine the overlap of images in image sets was presented.

(Pollefeys et al., 2008) have built a system that has been used to reconstruct 3D structure from sequences with more than one hundred thousand images. (Agarwal et al., 2009) and (Frahm et al., 2010) have presented approaches which can deal with hundred thousands or even millions of images from Community Photo Collections from the Internet to model urban areas. A major difference between these two approaches is, that the former runs on a cloud, the latter on just one multi-GPU PC system. While both approaches are impressive, one has to note that they are based on certain characteristics of the data and a couple of assumptions which make them tractable:

- Images at tourist attractions are often taken from nearly the same spot and thus look alike. I.e., many similar images can be found even for extremely down scaled versions of the images.
- The goal is to reconstruct the obvious 3D structure, leading to impressive 3D reconstructions of highlights, such as the Colosseum in Rome. Yet, there might be images, possibly with wider baselines, that could be used to extend the geometrical coverage or even link the tourist attractions. This is not done, as it would mean a detailed comparison of many more images.

Opposed to the above approaches, our goal is the detailed 3D modeling of urban areas from high resolution wide-baseline image sets.

In Section 2 we present the methods for point detection, matching and robust parameter estimation, that we have improved and combined for orientation of possibly unordered image sets.

For dense reconstruction, the results of our orientation procedure are used as input for Semiglobal Matching – SGM (Hirschmüller, 2008), which we are about to extend to 3D (Section 3). Although SGM has been developed for short baseline image sets, we found that due to our precise relative orientation, very good depth estimates were also possible for wider baselines.

Results are presented in Section 4. We have processed several hundred images acquired from (micro) Unmanned Aircraft Systems (UASs) and obtained a detailed 2.5D model from aerial images of an urban area. For a combination of aerial images with terrestrial images we have generated a preliminary dense 3D reconstruction of a building comprising the roof as well as the facades. In addition, we present a preliminary result for dense 3D surface reconstruction from terrestrial images only. Finally, in Section 5 conclusions are given and future work is discussed.

2 ORIENTATION OF UNORDERED IMAGE SETS

Albeit (Lowe, 2004) presents with SIFT a powerful solution for the estimation of point correspondences for short-baselines,

reliable point-matching for wide-baseline images can be much harder. Therefore, there is a strong need for improved matching methods. In the following we describe our approach which is based on scale invariant point matching, least squares matching and robust bundle adjustment (Figure 1).

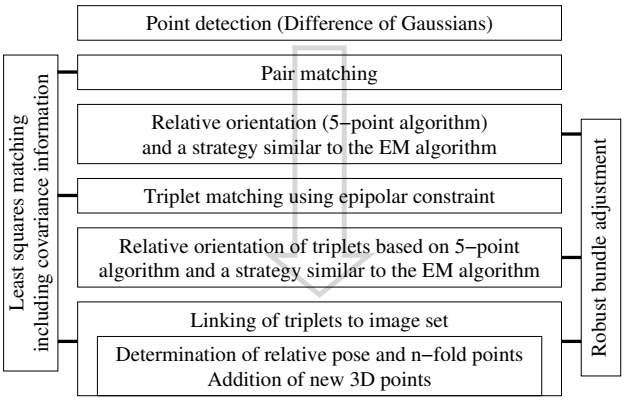

Figure 1: Image orientation based on scale invariant point matching, least squares matching and robust bundle adjustment.

Our approach for point matching produces reliable results even for major scale differences and illumination changes. It is based on normalized cross correlation (NCC), which is highly invariant against the latter, but only weakly against the former. To deal with scale differences, we use the concept of (Lindeberg, 1994) and determine points as scale space maxima based on Differences of Gaussians (DOG). NCC is weak against scale changes, because in this case image patches with the same size around conjugate points contain different scene parts. With the information obtained from scale space maxima, down-sampling of image patches with higher resolution is possible, so that they match to the same scene part. Potential point-matches obtained by scale-invariant NCC are refined by least squares matching (Grün, 1985) using an affine geometric model. This results into relative sub-pixel point positions including covariance information. The points and their covariance information are employed for relative orientation of pairs and triplets.

With the five point algorithm (Nistér, 2004) it became feasible to directly compute the relative orientation from calibrated image pairs. We have embedded a version of it into RANdom SAmple Consensus – RANSAC (Fischler and Bolles, 1981) using the Geometric Robust Information Criterion – GRIC (Torr and Zisserman, 1997). Additionally, a strategy similar to the Expectation Maximization (EM) algorithm is used to extend partial solutions.

Triplets and associated calibrated trifocal tensors are the basic geometric building block of our approach due to the following reasons:

- Opposed to pairs, where points can only be checked in one dimension by means of their distance from their respective epipolar lines, triplets allow for an unambiguous geometric checking of points. This does not only lead to much more reliable points, but also to improved, more reliable information for the cameras.
- Triplets can be directly linked into larger sets by determining their relative pose (translation, rotation and scale) from two common images.

Based on the highly reliable relative orientation, we are able to calculate the absolute orientation from unreliable and imprecise GPS data of low cost sensors, e.g., in a GPS camera, also in areas with strong occlusions, such as cities, similar to (Strecha et al., 2010).

To deal with unordered image sets, we employ an approach for automatic overlap detection consisting of the following steps:

- Determination of correspondences between the images
- Construction of two-view image matching graph
- Construction of three-view image matching graph

We use a fast GPU implementation (Wu, 2007) of SIFT to detect points and then match images pairwise, determining correspondences and hence the similarity between images. Available GPS information in the Exif tags of the images is used to derive the distance between images and thus to sort out unlikely pairs. This is employed to reduce the complexity. The resulting matching graph consists of images as nodes and edges which connect similar images. The weight of an edge is given by the number of correspondences between the two images, which is assumed to correspond to image similarity. Image pairs with the number of correspondences below a threshold will be considered as dissimilar and no edge is inserted.

Once the similarities between the images have been derived, we obtain a connected image set by constructing the maximum spanning tree (MST) of the matching graph using the modified algorithm of (Prim, 1957).

Finally, we determine triplets by iterating through the MST using the depth-first traversal algorithm. We discard triplets for which the images have a number of correspondences or a normalized overlap area below a threshold. For the determination of the overlap area between the images, we calculate the convex hull of the conjugate points of the triplet using the algorithm of (Sklansky, 1982). The area inside the convex hull is normalized by division through the total image area.

Our state concerning unordered image sets is still preliminary. Many image pairs, which can be oriented by our robust matching method, are not found due to the limited capability of the employed fast matching method (Wu, 2007). Therefore, while a combination of aerial and ground images is possible, the detection of overlapping images has to be conducted manually until a fast implementation of the robust point-matching method for wide-baseline images is available.

3 DENSE RECONSTRUCTION

For dense reconstruction we employ the original implementation of Semiglobal Matching – SGM (Hirschmüller, 2008). It is based on

- mutual information (MI) or the Census filter for cost computation and
- the substitution of a 2D smoothness term by a combination of 1D constraints (semiglobal).

MI presents the conditional probability distribution for the intensities in the matching image given an intensity in the reference image without resorting to a parametric model. Thus, MI can compensate a large class of global radiometric differences.

Though, one has to note that the conditional probability is computed for the whole image. This can be a problem for local radiometric changes, e.g., if materials with very different reflection characteristics exist in the scene or lighting conditions change.

The Census filter was found to be the most robust variant for matching cost computation (Hirschmüller and Scharstein, 2009). It defines a bit string with each bit corresponding to a pixel in the local neighborhood of a given pixel. A bit is set if the intensity is lower than that of the given pixel. Census thus encodes the spatial neighborhood structure. A 7×9 neighborhood can be encoded in a 64 bit integer. Matching is conducted via computing the Hamming distance between corresponding bit strings.

The smoothness term of SGM punishes changes of neighboring disparities (operator $T[]$ is 1 if its argument is true and 0 otherwise):

$$E(D) = \sum_{\mathbf{p}} \left(C(\mathbf{p}, D_{\mathbf{p}}) + \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_1 T[|D_{\mathbf{p}} - D_{\mathbf{q}}| = 1] + \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_2 T[|D_{\mathbf{p}} - D_{\mathbf{q}}| > 1] \right) \quad (1)$$

- The first term consists of pixel matching costs for all disparities of D.
- The second term adds a constant penalty P_1 for all pixels $\mathbf{q}$ from the neighborhood $N_{\mathbf{p}}$ of $\mathbf{p}$, for which the disparity changes only slightly (1 pixel).
- The third term adds a larger constant penalty P_2 for bigger changes of the disparities. Because it is independent of the size of the disparities, it preserves discontinuities.
- As discontinuities in disparity are often visible as intensity changes, P_2 is calculated depending on the intensity gradient in the reference image (with $P_2 \geq P_1$).

In 2D, global minimization is NP hard for many discontinuity preserving energies $E(D)$. In 1D, minimization can be done in polynomial time via dynamical programing, which is usually applied within image lines. Unfortunately, because the solutions for neighboring lines are computed independently, this can lead to streaking. For the semiglobal solution, 1D matching costs are computed in different (practically 8) directions which are aggregated without weighting. In the reference image, straight lines are employed, which are deformed in the matching image.

By computing the disparity images D for exchanged reference and matching image one can infer occlusions or matching errors by means of a consistency check. If more than one pair with the same reference image is matched, the consistency check is conducted for all pairs only once.

With the above methodology, very dense disparities can be computed. Using the camera calibration, all points can be projected into 3D space leading to dense 3D point clouds. While the original work of (Hirschmüller, 2008) has shown how to derive 2.5D surface models, work on the derivation of 3D surfaces by means of triangulation of the 3D points dealing also with outliers has been started only recently.

To model large-scale scenes in full 3D, which can produce billions of 3D points, efficient processing with regard to the computational and memory costs is a must. For this, octrees were found to be very suitable. Hence, we use for meshing a triangulation based on balanced octrees introduced in (Bodenmüller, 2009). Besides removing redundancy, octrees are particularly useful for visibility-checks in multiple-view geometry.

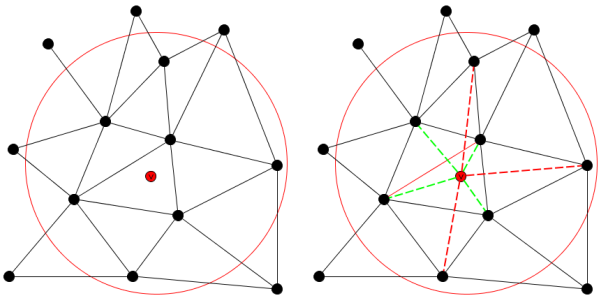

Figure 2: Local update of triangulation by adding a new vertex to temporary mesh. Left: New vertex v and projected neighborhood. Right (circle): New candidate edges (dashed lines). Green lines were accepted and red lines were removed, because of intersection with shorter edges.

Previous to mesh generation, points are eliminated depending on the angle of their normal vectors to the normal vectors of neighboring points. The triangle mesh is built incrementally. Iterating through all remaining points, the temporary mesh is projected on the tangent plane in a neighborhood of a new point (Figure 2). The new point is connected with all vertices within the neighborhood. If a new edge intersects an old edge in the plane, the longer one is removed.

After sketching our approach for dense 3D reconstruction, we present results for 2.5D and 3D surface reconstruction in the next section.

4 RESULTS AND DISCUSSION

We have applied our approach for orientation and dense reconstruction to several wide baseline image sets. For SGM in all cases the Census filter (Section 3) was used for cost computation.

The result in Figure 3 is based on 166 aerial images, acquired by a micro Unmanned Aircraft System (UAS). Although the flight was controlled automatically, the obtained image set is not very well structured. Because of too small overlap, many triplets cannot be matched. For this image set, we have qualitatively compared our approach with Bundler (Snavely, 2010). Particularly, we found that the relative orientation produced by Bundler is not very precise and the 3D point cloud contains many obviously false points. Thus, SGM could only be applied meaningfully after down-sampling the images to half the original resolution. In contrast, the relative orientation obtained by our approach is much more precise and could be used as basis for SGM on the original resolution images, leading to a much more detailed and realistic 2.5D model.

In Figure 4 we present a 2.5D model of an urban area obtained from aerial images acquired by a micro UAS from about 50 meters above the ground. The area with a size of about 600×100 meters was modeled from 262 images. Although the flight altitude was quite low, a detailed facade reconstruction was barely possible, because of the image configuration in combination with only 2.5D modeling. The roof overhang and the windows behind the facade render a realistic facade reconstruction not feasible in this case. For a complete and detailed 3D reconstruction of urban

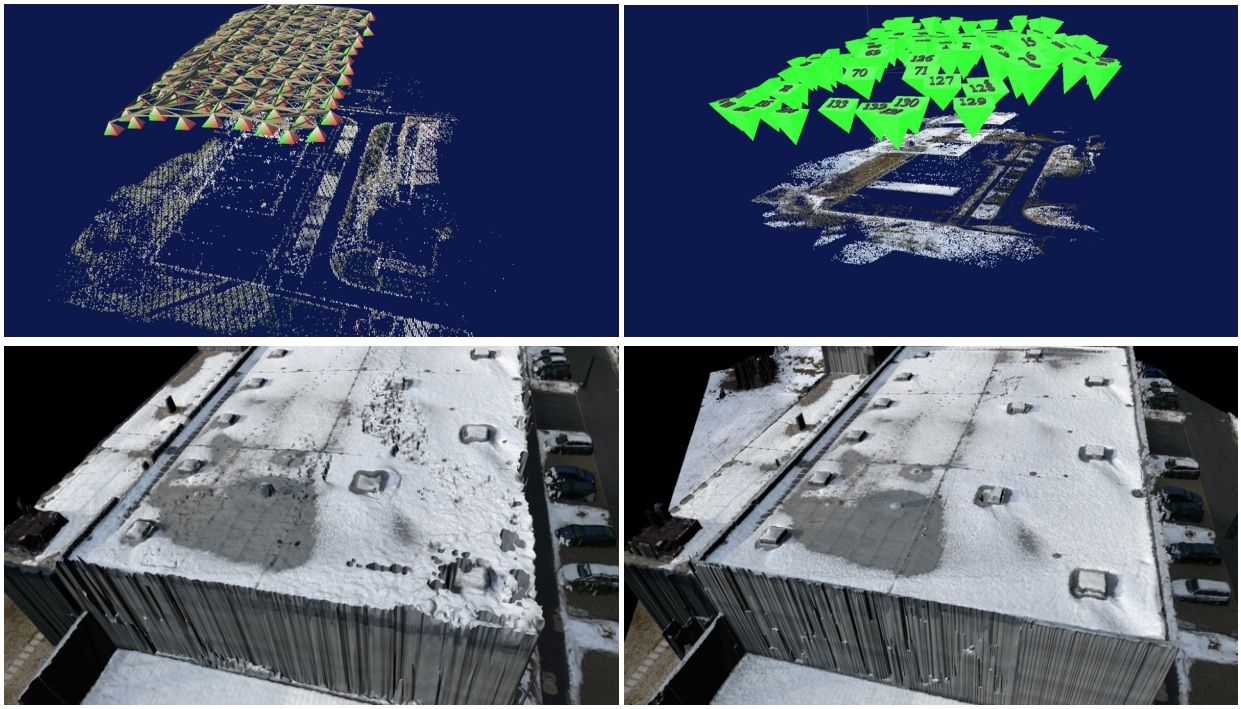

Figure 3: Top: Two relative orientations of a set of 166 images of a large Building in Wessling (Germany). Left: Relative orientation estimated by Bundler (Snavely, 2010) . Right: Relative orientation derived by our approach. Bottom: Resulting 2.5D models. Left: Result based on the orientation of Bundler. Right: Result using our orientation. The more precise relative orientation results in a much more detailed scene reconstruction.

areas there is, thus, a need for full, high resolution 3D reconstruction, that maintains the details and can deal with a combination of aerial and terrestrial images.

Figure 5 presents a preliminary result for our new approach for 3D surface reconstruction based on a combination of aerial and terrestrial images. The set of 205 images contains three different sequences:

- The flight sequence, which was taken from about 20 meters above the ground. It starts at the border of the village and covers buildings and terrain over several hundred meters.
- The terrestrial image-sequence acquired around one building, visible in the flight sequence.
- The "ascending" sequence connects aerial and terrestrial images. The images change in small steps from the ground to the bird's eye perspective. This image configuration is only feasible for micro UAS, which can be flown very close to facades and roofs.

The combination of the flight and the "ascending" sequence is quite difficult, because of major scale differences in combination with small overlap and perspective distortion. The dense 3D point cloud from SGM (Figure 5, second image) was generated with images from the terrestrial and the flight sequence and gives an almost complete result for the building. It illustrates, that roof and walls exactly fit to each other and, thus, that our relative orientation is very precise (please particularly note the roof overhang). The textured 3D result (Figure 5, bottom) shows the roof textured with different colors. This is caused by the different lighting conditions during the acquisition of the ground and the flight sequences.

Figure 6 shows another preliminary result of our work for full dense 3D surface reconstruction. We have applied our approach to the image sets "fountain-R25" and "castle-R20" of (Strecha et al., 2008). Due to our robust matching approach, a combination of both sets was possible.

5 CONCLUSIONS AND FUTURE WORK

In this paper we have presented an approach for automatic orientation and dense reconstruction from wide baseline image sets. As key characteristics it aims at a high precision in every step of the approach from least squares matching to robust bundle adjustment. Currently, our approach for full, high resolution 3D reconstruction does not maintain all the details that are available in the depth images of Semiglobal Matching. An optimal combination of depth estimation and smoothness priors has still to be investigated, but from our point of view the ability for precise relative orientation is fundamental.

Although our point matching approach is pretty robust against scale and illumination changes, it is not robust enough concerning viewpoint changes. Currently, arguably the best known concept for matching robust concerning viewpoint changes is the approach of (Morel and Yu, 2009). It simulates off-image-plane rotations, but has not been integrated into an approach for 3D reconstruction. The improvement of our approach by a similar procedure is part of future work.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of Florian Seibel, Patrick Reidelstürz and Peter Stütz during the acquisition of the aerial images.

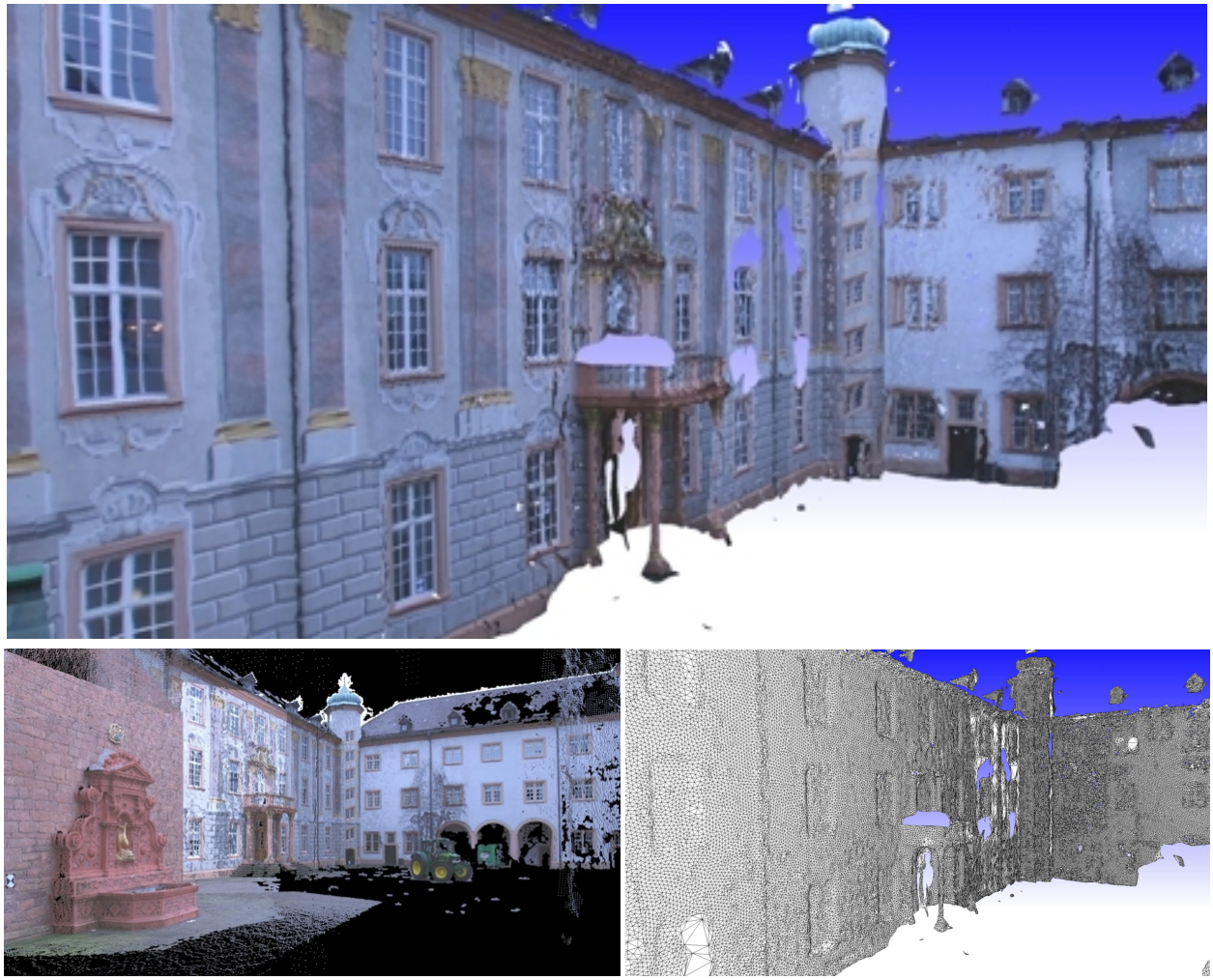

Figure 6: Top: Preliminary result for 3D reconstruction of a (textured) facade for Ettlingen Castle (Germany). Bottom: Left: Dense 3D point cloud. Right: Shaded 3D surface with wire-frames. The results are based on the image-sets "fountain-R25" and "castle-R20" of (Strecha et al., 2008).

REFERENCES

Agarwal, S., Snavely, N., Simon, I., Seitz, S. M. and Szeliski, R., 2009. Building Rome in a Day. In: 12th IEEE International Conference on Computer Vision (ICCV'09), pp. 72–79.

Bodenmüller, T., 2009. Streaming Surface Reconstruction from Real Time 3D Measurements. PhD Dissertation, Technichal University Munich, Germany.

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), pp. 381–395.

Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S. and Pollefeys, M., 2010. Building Rome on a Cloudless Day. In: 11th European conference on Computer vision: Part IV, ECCV'10, Berlin, Heidelberg, pp. 368–381.

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.

Hirschmüller, H. and Scharstein, D., 2009. Evaluation of Stereo Matching Costs on Images with Radiometric Differences. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(9), pp. 1582–1599.

Lindeberg, T., 1994. Scale-Space Theory in Computer Vision. Kluwer Academic Publishers, Boston, USA.

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

Morel, J. and Yu, G., 2009. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. SIAM Journal on Imaging Sciences 2(2), pp. 438–469.

Nistér, D., 2004. An Efficient Solution to the Five-Point Relative Pose Problem. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(6), pp. 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. and Towles, H., 2008. Detailed Real-Time Urban 3D Reconstruction from Video. International Journal of Computer Vision 78(2–3), pp. 143–167.

Figure 4: 2.5D Model of an urban area near Hammelburg (Germany). Top: Relative orientation including camera positions (red pyramids) and 3D points. Center: Overview of the 2.5D model. Bottom: Two different views from the ground.

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

Pollefeys, M., Vergauwen, M. and Van Gool, L., 2000. Automatic 3D Modeling from Image Sequences. In: International Archives of Photogrammetry and Remote Sensing, Vol. (33) B5/2, pp. 619–626.

Prim, R. C., 1957. Shortest Connection Networks and some Generalizations. Bell Systems Technical Journal pp. 1389–1401.

Schaffalitzky, F. and Zisserman, A., 2002. Multi-view Matching for Unordered Images Sets, or "How Do I Organize My Holiday Snaps?". In: Seventh European Conference on Computer Vision (ECCV'02), Vol. I, pp. 414–431.

Sklansky, J., 1982. Finding the Convex Hull of a Simple Polygon. Pattern Recognition Letters 1 pp. 79–83.

Snavely, N., 2010. Bundler: Structure from Motion for Unordered Image Collections. `http://phototour.cs.washington.edu/bundler/`.

Strecha, C., Pylvanainen, T. and Fua, P., 2010. Dynamic and Scalable Large Scale Image Reconstruction. In: 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10).

Strecha, C., Von Hansen, W., Van Gool, L., Fua, P. and Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. IEEE Conference on Computer Vision and Pattern Recognition (2008) pp. 1–8.

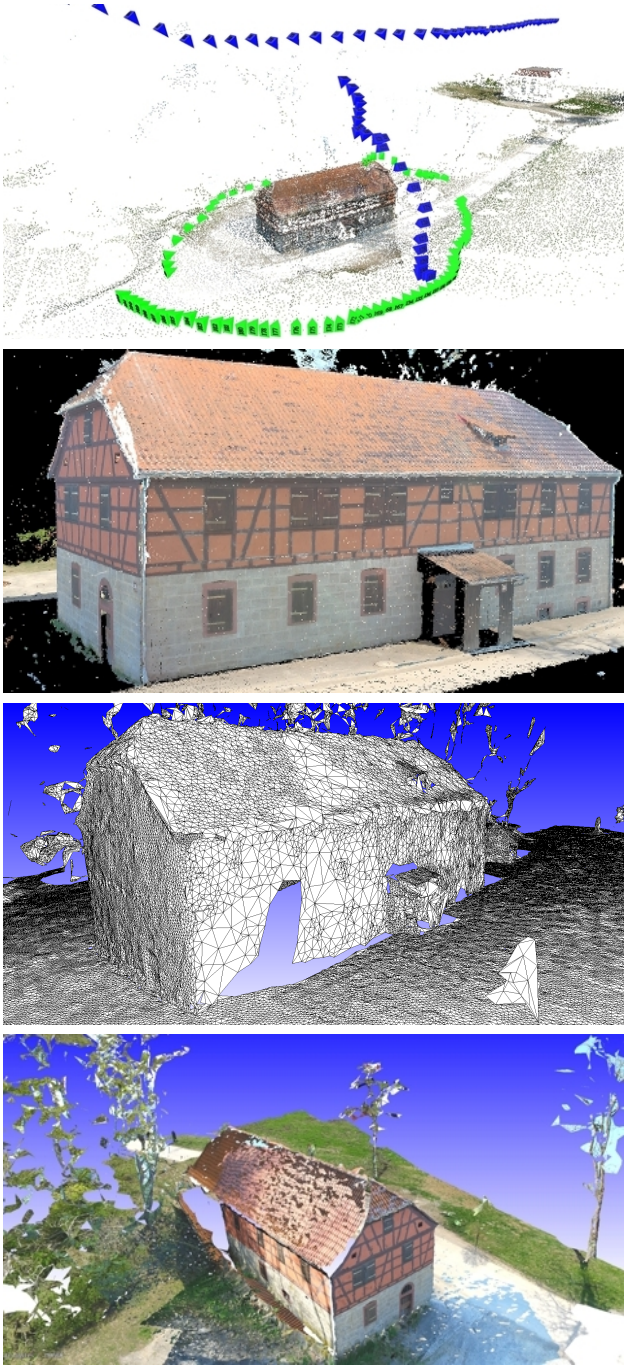

Figure 5: Top: Relative orientation of a combination of aerial (blue pyramids) and terrestrial (green pyramids) images of a building near Hammelburg (Germany), including camera positions and 3D points. Center: Dense 3D point cloud as result of relative orientation and Semiglobal Matching. Bottom (two images): Preliminary results for dense 3D surface reconstruction of the building, shaded and with wire-frames as well as textured. The combination of aerial and ground images renders it feasible to reconstruct the roof as well as the facades.

Torr, P. and Zisserman, A., 1997. Robust Parametrization and Computation of the Trifocal Tensor. Image and Vision Computing 15, pp. 591–605.

Wu, C., 2007. SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). `http://cs.unc.edu/~ccwu/siftgpu`.