Fast and Robust Generation of Semantic Urban Terrain Models from UAV Video Streams

Mathias Rothermel,
Norbert Haala, Konrad Wenzel
Institute for Photogrammetry
Stuttgart University, 70174 Stuttgart, Germany
forename.surname@ifp.uni-stuttgart.de

Dimitri Bulatov
Fraunhofer Institute of Optronics,
System Technologies and Image Exploitation
Gutleuthausstr. 1, 76275 Ettlingen, Germany
dimitri.bulatov@iosb.fraunhofer.de

***Abstract*—We present an algorithm for extracting Level of Detail 2 (LOD2) building models from video streams captured by Unmaned Aerial Vehicles (UAVs). Typically, such imagery is of limited radiometric quality but the surface is captured with large redundancy. The first contribution of this paper is a novel algorithm exploiting this redundancy for precise depth computation. This is realized by fusing consistent depth estimations across single stereo models and generating a 2.5D elevation map from the resulting point clouds. Disparity maps are derived by a coarse-to-fine Semi-Global-Matching (SGM) method performing well on noisy imagery. The second contribution concerns a challenging step of the context-based urban terrain modeling: Dominant planes extraction for building reconstruction. Because of noisy data and complicated roof structures, both dominant plane parameters and initial values for support sets of planes are obtained by the J-Linkage algorithm. An improved point-to-plane labeling is presented to encourage the assignment of proximate points to the same plane. This is accomplished by non-local, Markov Random Field (MRF) - based optimization and segmentation of color information. The potential and the limitations of the proposed methods are shown using an UAV video sequence of limited radiometric quality.**

I. Introduction

3D city models mainly consist of a digital elevation model and LOD2 building models showing detailed roof structures and planar façades [1]. In their capacity as pure geometry models, such 3D city models can be used as virtual environments for tourism or navigation systems, as well as for visualizations and training simulations. Common representations for surface geometry are digital elevation models or triangle meshes. However, in many applications, extensive simplification of this geometry is required to improve the interoperability. Additional context information, for example, differentiation between buildings and vegetation enables extraction of collision geometry for purposes of simulation.

3D City modeling long time was primarily based on LiDAR data due to its superior density and precision. Advances in sensor technology as well as algorithms for orientation computation and dense matching make image based reconstruction an efficient alternative. The potential of reconstructions using imagery captured by high quality, large-frame airborne cameras was demonstrated in [2], [3], [4]. State-of-the-art algorithms reconstruct surface points for almost each pixel and offer accurate results at depth discontinuities. For the mapping of moderate sized areas UAVs equipped with consumer grade cameras can be deployed [5], [6]. Low flying altitudes and velocities allow for data collection providing high detailed surface observations at high redundancy. However, in comparison to professional imaging devices imagery offers only limited signal-to-noise ratios.

In the first part of this article, we focus on the generation of 2D elevation data from highly redundant image data, however, suboptimal in their radiometry. In Sec. II, we describe a coarse-to-fine modification of SGM [2] which dynamically reduces search ranges for pixel correspondences. We show that by this modification, ambiguities in the correspondence problem are reduced and completeness of the reconstruction can be increased. To improve precision and eliminate gross errors, the high redundancy is exploited. Therefore, stereo matching is coupled with a correspondence linking strategy [7] where a set of stereo models is computed for each image. To refine results, redundant depth estimations are linked across the models, checked for consistency and fused subsequently. Moreover, a simple and efficient strategy for merging depth maps into 2.5D elevation models is presented. One major advantage of the proposed method is its scalability regarding resolution and the number of images. The increased computational load emerging from redundant processing can be tackled by parallel programming, reduction of search ranges including the efficient formulation for the 3D structure computation, thus allowing an efficient computing time of under 2 sec. per stereo model evaluation.

Within the second part of this paper, we focus on the extraction of building polyhedrons from the previously generated elevation maps. Thereby roof detail analysis, the vectorization of roof edges, ridges and smaller components as dormers is the most challenging step [8], [9], [10]. Our goal is to represent roof structure by as few polygons as possible while preserving a maximum amount of geometric information. Therefore, a common approach is to identify a set of dominant planes in the building point clouds. Global approaches, such as multi-model RANSAC [11] and J-linkage [12] operate on larger portions of points, for example point clouds representing complete buildings. These algorithms yield a set of plane hypotheses, each supported by a set of points. These support sets give valuable clues about spatial extend and impact of a plane hypothesis. However, the mentioned algorithms may extract erroneous planes (ghost planes) caused by support of points not representing the same geometric entity. This hinders a meaningful vectorization and can be overcome by forcing homogeneous labeling of neighbouring points [13], [9]. As will be explained in section III-B, one solution to this problem is a non-local, MRF - based optimization algorithm which results in a significant reduction of ghost planes. We will show that in presence of gross errors in the point clouds these results can be further refined by segmentation of color information of an ortho photo. In Sec. IV, we evaluate the presented algorithms for a UAV video sequence of a small village. Sec. V summarizes our contributions and outlines the ideas for future research.

II. Extraction of 2.5D Height Raster Data from Images and Videos

The input of the presented algorithm is a set of images and the corresponding interior and exterior orientations of cameras. These can be derived using structure from motion and bundle adjustment techniques such as [14] or [15]. Each available image $\mathbf{I}_b$ is treated as a reference image and is stereo matched against a set of match images $\mathbf{I}_{m,i}$. The selection of images corresponding to a cluster $(\mathbf{I}_b, \mathbf{I}_{m,i})$ is based on evaluation of camera poses assuring base lines and viewing directions to be in an appropriate range. To speed up the matching process, images are rectified to epipolar geometry.

In order to reject mismatches and improve accuracy, depth estimations in $\mathbf{I}_b$ are linked to correspondent depths in $\mathbf{I}_{m,i}$ and checked for geometric consistency. This results in a refined depth map or point cloud. We are aware of superior techniques evaluating image consistency across multiple views [16], [17], [18], however, to limit processing time we rely on stereo processing and subsequently filter mismatches based on geometric properties. In the final stage, all refined depth maps are fused using a gridding approach resulting in a 2.5D surface model.

A. Depth Map Generation and Refinement

The implemented stereo algorithm is a coarse-to-fine modification of the SGM method [2]. The problem of dense matching can be stated as densely estimating the correspondences $\mathbf{x}_{bi}$ and $\mathbf{x}_{mi} = \mathbf{x}_{mi}(d)$ across a reference view $\mathbf{I}_b$ and a second frame $\mathbf{I}_m$. In the first stage, a photo-consistency measure $E_{data}(d_\mathbf{x})$ is computed [19] and stored for each potential correspondence pair $(\mathbf{x}_b, \mathbf{x}_{mi})$. The search range for correspondences is given by the 1D horizontal epipolar line and a pre-defined, constant disparity search range. Secondly, costs of photo-consistency are accumulated along eight image paths in order to force piecewise smoothness of the underlying surface. The set of minimal accumulated costs yields a set of disparities $\mathbf{D}$ corresponding to a strong local minimum of an energy function:

$$E(\mathbf{D}) = \sum_{\mathbf{x}} E_{data}(d_\mathbf{x}) + \sum_{\mathbf{x},\mathbf{y} \in \mathcal{N}} E_{smooth}(d_\mathbf{x}, d_\mathbf{y}), \qquad (1)$$

where $\mathcal{N}$ represents the neighbourhood of a pixel, and E_{smooth} is usually the truncated linear penalty term

$$E_{smooth}(d_\mathbf{x}, d_\mathbf{y}) = \min\left(\lambda_1 \left|d_\mathbf{x} - d_\mathbf{y}\right|, \lambda_2\right), \; 0 < \lambda_1 \leq \lambda_2. \qquad (2)$$

In contrast to the original approach, we evaluate potential correspondences only in a reduced search range as visualized in Fig. 1. The modification aims at reducing ambiguities of photo consistency measures. This is particular beneficial for matching imagery of low radiometric quality since photo consistency measures in general are less distinctive. Within our approach the search range is limited for each pixel individually based on disparities available from disparity maps already computed on the next higher pyramid level. The choice of a suitable range is crucial to assure small details possessing large depth changes to be reconstructed completely. In our algorithm, ranges are derived by analysis of the local surface structure and validity in a window around the respective pixel. On the highest pyramid level, correspondences are searched along the full range of epipolar lines.

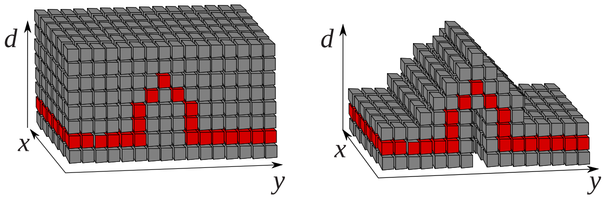

Fig. 1: Visualization of cost structures of classic SGM (left) and the dynamic solution (right). Red cubes represent costs for the true correspondences. Gray cubes mark the costs of potential correspondences.

The high redundancy in the image sequences is exploited by stereo matching each frame with multiple proximate frames to retrieve consistent depth estimations for each pixel. Thereby, multiple measurements across i disparity maps in one image cluster $(\mathbf{I}_b, \mathbf{I}_{m,i})$ are fused using the concept of correspondence linking inspired by [7]. A base image pixel $\mathbf{x}_b$ can be related to the pixel coordinates $\mathbf{x}_b^r$ in the rectified base images $\mathbf{I}_{b,i}^r$ by the homographies $\mathbf{H}_i$ used within the rectification process $\mathbf{x}_{b,i}^r = \mathbf{H}_i \mathbf{x}_b$. For the rectified coordinates, disparity estimates $d_i^r(\mathbf{x}_{b,i})$ can be derived by lookup in the disparity maps available from matching. The relation between the disparity d and the depth D of the corresponding 3D point with the z-coordinate Z in a relative coordinate system can be calculated using the formula for the stereo normal case [20]:

$$d = \frac{fB}{Z(\mathbf{x}_b^r)} = \frac{Ba}{D(\mathbf{x}_b)}, \text{ where } a = \sqrt{f^2 + (x_b^r)^2 + (y_b^r)^2}, \qquad (3)$$

and f is the focal length and B the baseline. The fact that distances between camera centers and object points remain the same in the original and the rectified coordinate systems implies that depths $D(\mathbf{x}_{b,i}^r)$ computed for single stereo models equal depths on the base image ray, thus $D(\mathbf{x}_{b,i}^r) = D_b$. We calculate the final depth from the set of i consistent disparities by minimizing the 1D reprojection error $|x_{m,i}^r - \hat{x}_{m,i}^r|_2$ along the epipolar lines. This is the same as claiming $\|d_i - \hat{d}_i\|_2 \overset{!}{=} \min$. Substitution of equation 3 and derivation leads to

$$D = \left(\sum_i B_i^2 a_i^2\right) \left(\sum_i B_i a_i d(\mathbf{x}_{b,i}^r)\right)^{-1}. \qquad (4)$$

In order to discard spurious disparities for the multi-view triangulation, we assume that each disparity d_i is estimated with a certain precision σ_i and calculate, by means of considerations on error propagation for 3, the uncertainty interval $D_i \pm \sigma_i$. If these ranges mutually overlap, respective disparities are considered consistent and used within the subsequent triangulation step. If the number of consistent measurements is below a threshold t_c, the point is not triangulated.

B. Depth Map Fusion for 2.5D Elevation Maps

For the application of simulation, we strive for a possibly complete (water-tight) representation of urban terrain, and in particular, buildings. Because it is not realistic to cover all walls of all buildings during a UAV flight, we merely want to reconstruct the roof structure of buildings and add wall polygons at roof edges at later stage. Within the present approach we fuse 3D clouds available from the reconstruction process into a 2.5D elevation map. Automatic correction of wall positions using the available façade points will be addressed in our future work. We assign all extracted points to a $n \times m$ grid located parallel to the ground. The dimension of a single grid cell corresponds to the average pixel footprint. Let p be the average number of points assigned to one grid cell. In order to preserve clear edges, points representing 3D structure are removed by only considering the p highest points. The final height value of each grid cell is then computed as the median of all z-components of the assigned points.

Additional to the height data, a color value is computed for each grid cell. The resulting RGB image will be referred to as true ortho photo. The color value for each cell is computed as average color of all assigned points. Its dimension and resolution are equal to that of the elevation map. It is used to support dominant plane extraction and to texture the final model.

III. Extraction of Semantic Models from Elevation Maps

City models contain geometric entities of different object types, such as buildings, trees, streets, etc. In this work, we are interested in the extraction of geometric primitives from point (sub-)clouds representing building roofs. Based on the generated elevation data, the set of points representing buildings is identified as the set of points neither classified as ground nor as vegetation. We derive ground points following [21]. For classification of vegetation points, the Normalized Difference Vegetation Index is applied using the green channel and the maximum of red and blue channels. The remaining points form our building hypothesis. Spurious points or point patches are filtered by means of minimal area, minimal eccentricity and local height differences [8]. Moreover the evaluation of local height differences allows a further subdivision of larger building complexes. Additional or alternative ways to reduce the set of points is to detect small roof segments (such as chimneys, see [22]) in advance and then subdivide building complexes along their diagonals (see [23]) once building ground polygon has been obtained. The obtained sub-clouds represent roof structure of buildings and comprise the main input for the subsequent vectorization.

A. Dominant Plane Computation

Assuming piecewise planarity of the roof structure, a set of planes can be fitted to a building point cloud. For this purpose, multi-model RANSAC is a well-known method that can be successfully applied at least for simple configurations [24]. Assessing the hypotheses of RANSAC according to the geometric distributions of inliers [25] or preferring neighbouring points for hypotheses generation make the global method of RANSAC more local and reduces the number of erroneous hypothesis, so-called ”ghost planes”. The latter strategy constitutes one significant idea of the J-linkage [12] algorithm which is used in this work. Another innovation of J-linkage is not to discard a hypothesis after a better one was found, but keeping a user-specified number of hypotheses. The extracted planes can be intersected to retrieve the desired cut lines. Besides dominant planes, the set of points supporting each hypotheses plays an essential role. These support sets hold information regarding spatial extend and impact of each plane hypotheses. For the point-to-plane assignment (labeling), different strategies can be used. Two examples are assigning the label of the first plane that has the point as inlier (finder) or assigning the label of the plane possessing the minimum distance (winner). However, no constraints regarding spatially homogeneous labeling is applied and assignments supporting erroneous plane hypotheses can not fully be avoided.

B. Point Labeling Using Non-Local Optimization and Segmentation

Given a set of planes, we want to impose a soft constraint such that neighbouring points $\mathbf{y}$ of $\mathbf{x}$ possess the same label $l_{\mathbf{x}}$. Therefore, we use discrete optimization and consider the same form of cost function as given in equation (1). Instead of optimizing the disparity of a pixel $d_{\mathbf{x}}$ as in Sec. II, the label $l_{\mathbf{x}}$ is optimized. $E_{data}(l_{\mathbf{x}})$ is the truncated and rescaled distance of $\mathbf{x}$ to the plane $\mathbf{p}_l$ labeled by l from and E_{smooth} is a special case of equation (2) for $\lambda_1 = \lambda_2 = \lambda$, which is the well-known Potts term multiplied by a constant:

$$\begin{aligned} E_{data}(l_{\mathbf{x}}) &= 2^{11} \min\left(\mathtt{dist}(\mathbf{x}, \mathbf{p}_l)/\delta, 1\right) \\ E_{smooth}(l_{\mathbf{x}}, l_{\mathbf{y}}) &= 0 \text{ if } l_{\mathbf{x}} = l_{\mathbf{y}} \text{ and } \lambda \text{ otherwise.} \end{aligned} \tag{5}$$

Thereby δ is the inlier threshold multiplied by a scalar (1–2). The normal vector is rescaled by one. The scale 2^{11} is needed in order to perform computations with 16 bit integers. For the smoothness parameter λ, a constant value of 1000 was used in this work.

The simplest non-local method in order to find an approximation for the solution of the Markov Random Field problem (equation (1)) is to apply the dynamic programming as proposed by [26]. This method only considers 1D neighborhoods and operates on every scanline. We do not go into the details of this method, but we note that it is a special case of SGM accumulating costs along one path only. The computation of the 1-D minimum of equation (1) is especially fast in case of the computational inexpensive smoothness function provided by equation (5). Since the data term in equation (2) is already very distinctive, the performance for different scan-lines is quite the same and the running time is below one second even for buildings containing around 50000 pixels and 15 labels.

For datasets with many outliers, the labeling of the non-local optimization method described above can be further refined by segmentation of the ortho photo, by means of [27]. We discard those segments that are to small and too narrow. A further assumptions is that if almost all pixels of the segment are part of the roof segment, the rest of its pixels will likely also belong to the same roof segment. For a pixel $\mathbf{x}$, we have

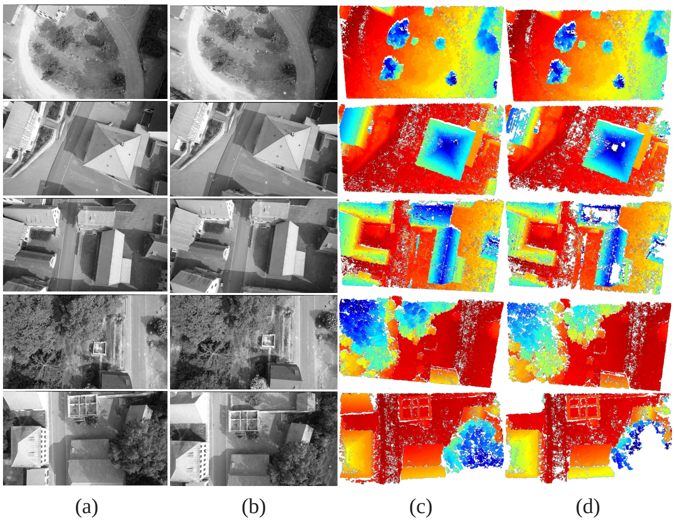

Fig. 2: Disparity maps generated from input images (a)(b). Our approach(c) outperforms the classical SGM (d) on repetitive structures (roofs) and low textured areas (streets).

a new label $\hat{l}$:

$$\hat{l}(\mathbf{x}) = \begin{cases} l & \text{if } |S(\mathbf{x}) \cap (=l)| > \gamma\,|S| \\ l(\mathbf{x}) & \text{otherwise,} \end{cases}$$

where S is the segment label and $\gamma \approx 0.8$ is a scalar. After extraction of dominant planes for one building and polygonizing the label masks, the neighbouring polygons are intersected in space between each other and building outlines, thus encouraging a consistent representations of building roofs.

IV. Results

We demonstrate the performance of the presented approach on a urban UAV video sequence captured over the village Bonnland in Southern Germany. The flight instability of UAVs result in considerable variations in camera orientation. Thus, captured imagery is rather unstructured and often blurred. To guarantee sufficient base lengths for the respective image pairs, every 20th frame was extracted from the sequence resulting in 647 images. Strucure from motion and bundle block adjustment were carried out using a state-of-the-art tool [15]. Self calibration was performed to estimate distortion and internal camera parameters.

Figure 2 shows five examples of stereo pairs and respective disparity maps generated by the proposed method and the classical SGM. Thereby all parameter settings were identical. It becomes clear that higher point density can be achieved using the dynamic approach both in weakly textured image areas (streets, shadows) and in areas with repetitive texture (roofs, vegetation). This is because ambiguities are reduced by narrowing down the search ranges. However, reconstruction of areas close to height jumps is more accurate for the classical SGM method. The depth maps by the coarse-to-fine solution were generated in the range from 1.9 to 2.4 seconds including I/O operations using a I7 quad-core processor. The maximum memory consumption was in the range of 0.1GB to 0.3GB. Disparity maps using SGM were constructed in 3.3 to 9.4 seconds using a maximum of 0.4GB to 1.6GB of RAM. However, our SGM program uses the same core implementation as the dynamic solution and could possibly be implemented in a more efficient way.

To demonstrate the surplus of the correspondence linking method, we computed point clouds using different thresholds t_c for the minimal number of spatially consistent depth estimations. For the center of the test area, all points clouds were merged in model space and the number n_e obvious blunders above the church top and below the ground was recorded and removed for visualization purposes. Fig. 3a depicts points generated from pure stereo matching ($t_c = 0$). Here, the result possesses a large number of blunders $n_e = 20141$ and is rather noisy. By claiming spacial consistency of $t_c = 2$ in Fig. 3b and $t_c = 3$ in Fig. 3c, the number of blunders becomes significantly lower: $n_e = 18$ and $n_e = 0$, respectively, and the noise level is drastically reduced. As shown in Fig.3a, 3c, and 3c large parts of the building walls are reconstructed. The point cloud representing the elevation map is depicted in Fig. 3d. Thereby results from correspondence linking with $t_c = 2$ were fused as explained in section II-B. It can be observed that 3D structure at house walls is reliably removed, roof edges are extracted precisely, the noise level is further reduced while small details as chimneys or dormers are preserved.

For the evaluation of J-linkage coupled with the improved point labeling, we concentrate on a rather challenging building in the test area, depicted in Fig. 4. Input points are colored black. Fig. 4a visualizes the point labeling for RANSAC using the winner strategy. Fig. 4b depicts results of J-Linkage using the same strategy. RANSAC yields one ghost plane (specified by cyan colour, arrow 1) while one big plane is lost (arrow 2). This plane is detected by J-linkage (dark-green). Due to inlier thresholds, both procedures are not able to detect dormers (arrow 3) and other smaller planes; efforts will be made in the future to identify these segments either before or after the dominant planes computation process. By using J-linkage and the finder strategy (Fig. 4c) the number of ghost planes is reduced. However, for all approaches narrow segments remain, which degrade the process of polygonizaition of roof details. Mostly, these points are inliers of more than one plane but are assigned to a plane with majority of inliers in another part of the building. By applying J-linkage and the subsequent discrete optimization algorithm proposed in III-B (Fig. 4d), ghost planes can be further reduced. The small remaining mislabelings – for example for points representing the power transmission lines (arrow 4) – can be removed by segmentation of an orthophoto as explained in III-B.

The complete reconstruction based on the point cloud derived by all 647 UAV-video frames is shown in Fig. 5. Beside the video stream for reconstruction of geometry and texturing roofs, several oblique videos were registered into the model coordinate system and used for texturing building walls. Fig. 6 shows three examples of building models.

V. Conclusions and Outlook

We presented an image-based reconstruction method handling images sets with challenging radiometric and geometric properties of a video stream captured by an UAV. It is based on a coarse-to fine modification of the SGM algorithm with dynamically adapted disparity search ranges. The method

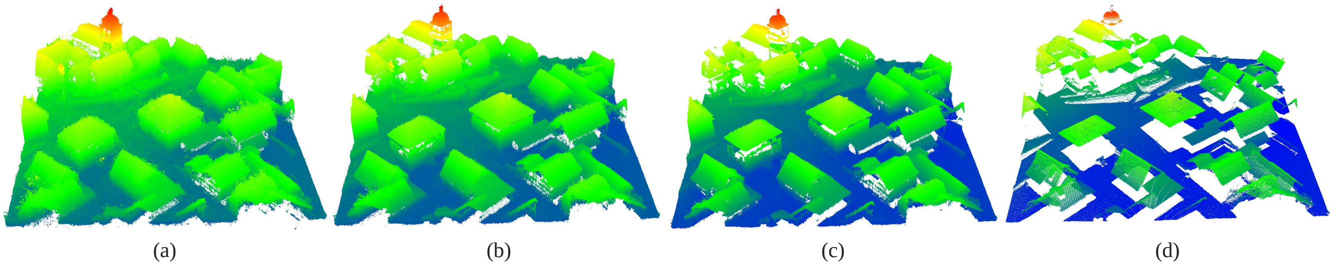

Fig. 3: Point clouds for test area depending on different number of minimal geometric consistent disparities t_c. See text for more comments.

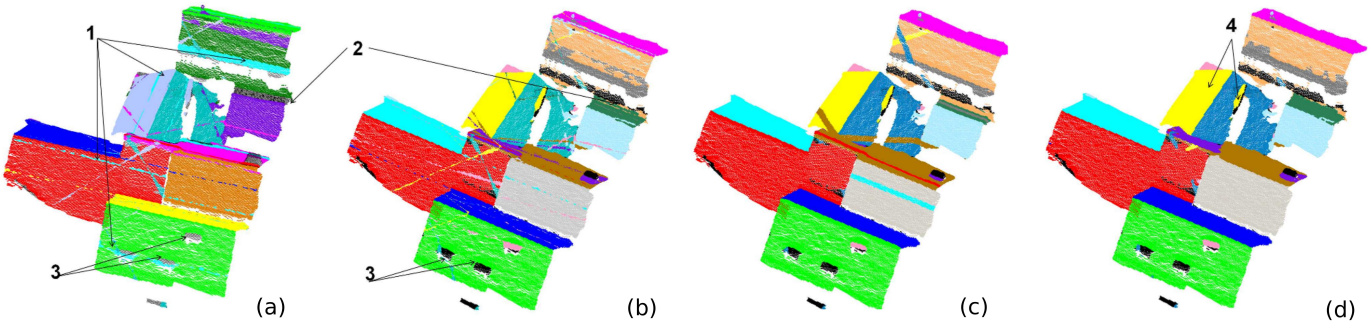

Fig. 4: Performance of RANSAC and J-Linkage algorithm for a complicated building. See text for more comments.

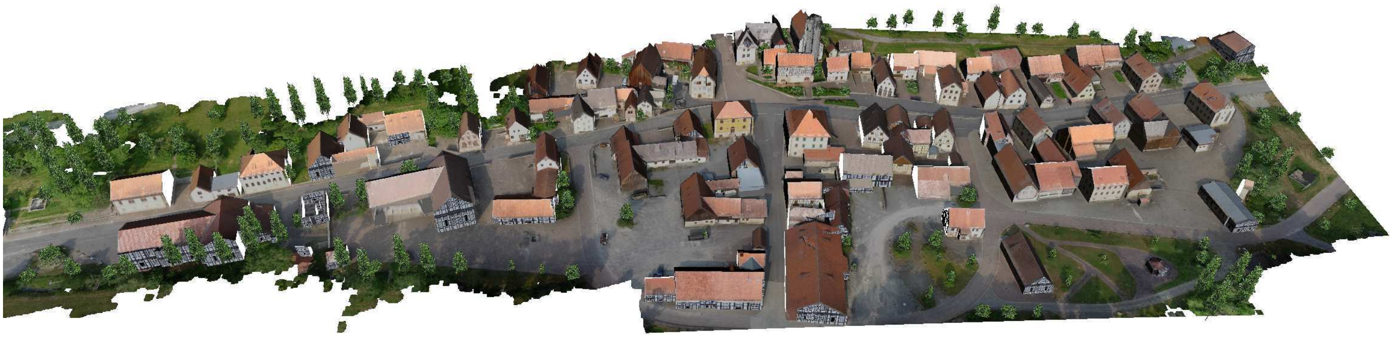

Fig. 5: A view of the textured reconstruction of the data-set Bonnland by our algorithm.

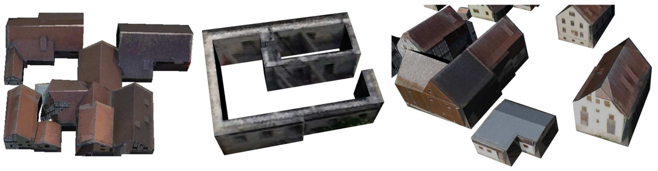

Fig. 6: Three exemplary building models of higher complexity.

allows to reduce both computation time and memory load for stereo-pair disparity computation. Moreover, higher densities of reconstructions for image parts of repetitive texture and low signal-to-noise ratios were achieved, however, with a slight reduction of accuracy in regions with jumps of elevation. In order to exploit redundancy, stereo matching is coupled with a correspondence linking approach efficiently filtering blunders and increasing precision. We presented an algorithm for fusing these results by the generation of 2.5D elevation models. Although the elevation models are of impressive precision, real 3D structure cannot be adequately represented. This issue will be addressed in our future work aiming for the reconstruction of façades including entities such as windows and doors for LOD3 modeling purposes. The reconstruction pipeline is free for academic use and is publicly available at *http://www.ifp.uni-stuttgart.de/publications/software/*.

The standard building reconstruction procedure has its main innovation in the dominant plane extraction step. It is carried out by J-Linkage followed by a non-local optimization for point labeling which reliably reduces the influence of ghost planes and at the same time preserves small details.

Thereby elevation data can be compressed to several hundreds of polygons. For our future work towards real 3D geometry, we strive for improving positions of building walls by means of depth maps, foreground extraction, cleaning textures in front of the building walls, and annotation of the foreground objects. The optimization module will be replaced by graph-based methods (Alpha-Expansion and Alpha-Beta swap) which are excellent tools for minimization of cost functions on arbitrary graphs.

REFERENCES

[1] T. H. Kolbe, "Representing and exchanging 3d city models with citygml," in *3D geo-information sciences*. Springer, 2009, pp. 15–31.

[2] H. Hirschmüller, "Stereo processing by semi-global matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[3] H.-H. Vu, R. Keriven, P. Labatut, and J.-P. Pons, "Towards high-resolution large-scale multi-view stereo," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1430–1437.

[4] N. Haala and M. Rothermel, "Dense multi-stereo matching for high quality digital elevation models," *PFG Photogrammetrie, Fernerkundung, Geoinformation*, vol. 2012, no. 4, pp. 331–343, 08 2012. [Online]. Available: http://dx.doi.org/10.1127/1432-8364/2012/0121

[5] N. Haala and M. Rothermel, "Dense multiple stereo matching of highly overlapping uav imagery," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXIX-B1, pp. 387–392, 2012. [Online]. Available: http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XXXIX-B1/387/2012/

[6] A. Irschara, M. Rumpler, P. Meixner, T. Pock, and H. Bischof, "Efficient and globally optimal multi view dense matching for aerial images," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. I-3, pp. 227–232, 2012. [Online]. Available: http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/I-3/227/2012/

[7] R. Koch, M. Pollefeys, and L. J. V. Gool, "Multi viewpoint stereo from uncalibrated video sequences," in *Proceedings of the 5th European Conference on Computer Vision-Volume I - Volume I*, ser. ECCV '98. London, UK, UK: Springer-Verlag, 1998, pp. 55–71. [Online]. Available: http://dl.acm.org/citation.cfm?id=645311.649088

[8] H. Gross, U. Thönnessen, and W. v. Hansen, "3D-Modeling of urban structures," in *Proc. of Joint Workshop of ISPRS/DAGM Object Extraction for 3D City Models, Road Databases, and Traffic Monitoring CMRT05, International Archives of Photogrammetry and Remote Sensing*, vol. 36, Part 3W24, 2005, pp. 137–142.

[9] F. Lafarge and C. Mallet, "Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation," *International journal of computer vision*, vol. 99, no. 1, pp. 69–85, 2012.

[10] G. Sohn, Y. Jwa, H. B. Kim, and J. Jung, "An Implicit Regularization for 3D Building Rooftop Modeling using Airborne LIDAR Data," in *Proc. of ISPRS-Congress, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. I-3, 2012, pp. 305–310.

[11] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM (45)*, pp. 381–395, 1981.

[12] R. Toldo and A. Fusiello, "Robust multiple structures estimation with J-Linkage," in *Proc. of European Conference on Computer Vision (1)*, 2008, pp. 537–547.

[13] L. Zebedin, J. Bauer, K. F. Karner, and H. Bischof, "Fusion of feature- and area-based information for urban buildings modeling from aerial imagery," *ECCV*, vol. 4, pp. 873–886, 2008.

[14] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3d," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, Jul. 2006. [Online]. Available: http://doi.acm.org/10.1145/1141911.1141964

[15] G. Verhoeven, "Taking computer vision aloft–archaeological three-dimensional reconstructions from aerial photographs with photoscan," *Archaeological Prospection*, vol. 18, no. 1, pp. 67–73, 2011.

[16] D. Bulatov, P. Wernerus, and C. Heipke, "Multi-view dense matching supported by triangular meshes," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 6, pp. 907–918, 2011.

[17] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*. IEEE, 1996, pp. 358–363.

[18] E. Baltsavias, *Multiphoto Geometrically Constrained Matching*, ser. Institut für Geodäsie und Photogrammetrie Zürich: Mitteilungen. ETH, 1991. [Online]. Available: http://books.google.de/books?id=S1cnKQEACAAJ

[19] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European Conference of Computer Vision (ECCV 1994)*, 1994, pp. 151–158.

[20] K. Kraus, *Photogrammetrie - Band 1*. Ferd. Dümmlers Verlag, ISBN: 3-427-78645-5, 1994.

[21] D. Bulatov, P. Wernerus, and H. Gross, "On applications of sequential multi-view dense reconstruction from aerial images," in *Proc. of International Conference on Pattern Recognition Applications and Methods (2)*, 2012, pp. 275–280.

[22] D. Bulatov and M. Pohl, "Detection of small roof details in image sequences," in *Proc. of Scandinavian Conference on Image Analysis*, 2013, pp. 601–610.

[23] N. Haala, C. Brenner, and K.-H. Anders, "3D Urban GIS from Laser Altimeter and 2D Map Data," *ISPRS Congress Commission III, Working Group 4*, vol. 32, no. 3/1, pp. 339–346, 1998.

[24] D. Bulatov, F. Rottensteiner, and K. Schulz, "Context-based urban terrain reconstruction from images and videos," in *Proc. of ISPRS-Congress, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. I-3, 2012, pp. 185–190.

[25] F. Tarsha-Kurdi, T. Landes, and P. Grussenmeyer, "Extended ransac algorithm for automatic detection of building roof planes from lidar data," *The photogrammetric journal of Finland*, vol. 21, no. 1, pp. 97–109, 2008.

[26] P. Belhumeur, "A Bayesian Approach to Binocular Stereopsis," *International Journal of Computer Vision*, vol. 19(3), pp. 237–260, 1996.

[27] J. Wassenberg, W. Middelmann, and P. Sanders, "An efficient parallel algorithm for graph-based image segmentation," in *Proc. of Computer Analysis on Images and Patterns*, 2009, pp. 1003–1010.