Proceedings of the OAGM&ARW Joint Workshop 2017 DOI: 10.3217/978-3-85125-524-9-28

Fusion of Point Clouds derived from Aerial Images

Andreas Schönfelder[1,2], Roland Perko[1], Karlheinz Gutjahr[1], and Mathias Schardt[2]

Abstract— **State of the art dense image matching in combination with advances in camera technology enables the reconstruction of scenes in a novel high spatial resolution and offers new mapping potential. This work presents a strategy for fusing highly redundant disparity maps by applying a local filtering method to a set of classified and oriented 3D point clouds. The information obtained from stereo matching is enhanced by computing a set of normal maps and by classifying the disparity maps in quality classes based on total variation. With this information given, a filtering method is applied that fuses the oriented point clouds along the surface normals of the 3D geometry. The proposed fusion strategy aims at the reduction of point cloud artifacts while generating a non-redundant surface representation, which prioritize high quality disparities. The potential of the fusion method is evaluated based on airborne imagery (oblique and nadir) by using reference data from terrestrial laser scanners.**

I. INTRODUCTION

While the processing of aerial and satellite imagery for the generation of 2.5D Digital Elevation Models (DEM) from Multi-View Stereo (MVS) systems is a standard procedure in the field of photogrammetry and remote sensing, the reconstruction of complex 3D scenes poses several new challenges. Therefore, this work focuses on a 3D fusion of point clouds, in contrast to classical mapping approaches that only produce and fuse 2.5D DEMs or elevation maps (cf. [14]). In order to process large frame airborne and satellite imagery, it is necessary to ensure that the MVS system is capable of processing data of arbitrary size in adequate runtime at highest possible geometric accuracy.

The main contribution of this work is an easy to implement, scalable 3D point cloud fusion strategy which builds on classic multi-view stereo pipelines. By restricting, respectively weighting, disparities based on their quality it is possible to generate surface representations of large-scale datasets in adequate runtime, simultaneously reducing the redundancy in the point cloud and increasing the geometric accuracy.

II. STATE OF THE ART

Typically, the processing of multiple stereo images yields one depth map or disparity map per stereo pair. To generate one consistent, non-redundant representation of the mapped scene, the depth maps have to be fused. Some MVS systems tackle this problem by linking surface points directly in the process of image matching. In contrast, MVS systems like PMVS [4], use multi-photo consistency measures to optimize position and normals of surface patches and iteratively grow the surface starting from a set of feature points. In many MVS systems, depth maps are generated via Semi-Global Matching (SGM) [6] and spatial point intersection yielding one depth map per stereo pair. SGM is one of the most common stereo matching algorithms used in mapping applications offering robust and dense reconstruction while preserving disparities discontinues.

Depth map fusion or integration is one of the main challenges in MVS and different approaches have been developed over the last decades. Authors of [17] propose an excellent benchmark dataset for the evaluation of MVS surface reconstruction methods. As mentioned in [12], the Middlebury MVS benchmark test demonstrates that global methods tend to produce the best results regarding completeness and accuracy, while local methods like [3] offer good scalability at smaller computational costs. Moreover MVS methods can be categorized based on their representation which can differ from voxels, level-sets, polygon meshes up to depth maps [17]. Authors like [5] and [15] focus on the fusion of depth maps to generate oriented 3D point clouds. The surface reconstruction in terms of fitting a surface to the reconstructed and fused points is defined as a post-processing step which can be solved using algorithms like the generic Poisson surface reconstruction method proposed by Kazhdan *et al.* [8].

Regarding the processing of aerial imagery scalability is an important factor. As mentioned in [12], a number of scalable fusion methods have been presented in the last years, e.g. [3], [11], [18], yet they are still not able to process billions of 3D points in a single day or less [18]. Kuhn *et al.* [9] propose a fast fusion method via occupancy grids for semantic classification. The fusion method complements state-of-the-art depth map fusion as it is much faster. However, it is only suitable for applications that have no need for dense point clouds. All of the mentioned scalable fusion methods have in common, that octrees are used as underlying data structures. Kuhn *et al.* [10] introduce an algorithm for division of very large point clouds. They discuss different data structures and their capability for the decomposition of reconstruction space. In addition, Kuhn *et al.* [12] show that the 3D reconstruction of fused disparity maps can be improved by modeling the uncertainties of disparity maps. These uncertainties are modeled by introducing a feature based on Total Variation (TV) which allows pixel-wise classification of disparities into different error classes. Total variation in context with MVS was first introduced by Zach *et al.* [19]. They propose a novel range integration method using a global energy functional containing a TV regularization force and an L^1 data fidelity term for increased robustness to outliers.

[1]Joanneum Research Forschungsgesellschaft mbH, Steyrergasse 17, 8010 Graz, Austria {firstname.lastname}@joanneum.at

[2]Graz University of Technology, Steyrergasse 30, 8010 Graz, Austria {firstname.lastname}@tugraz.at

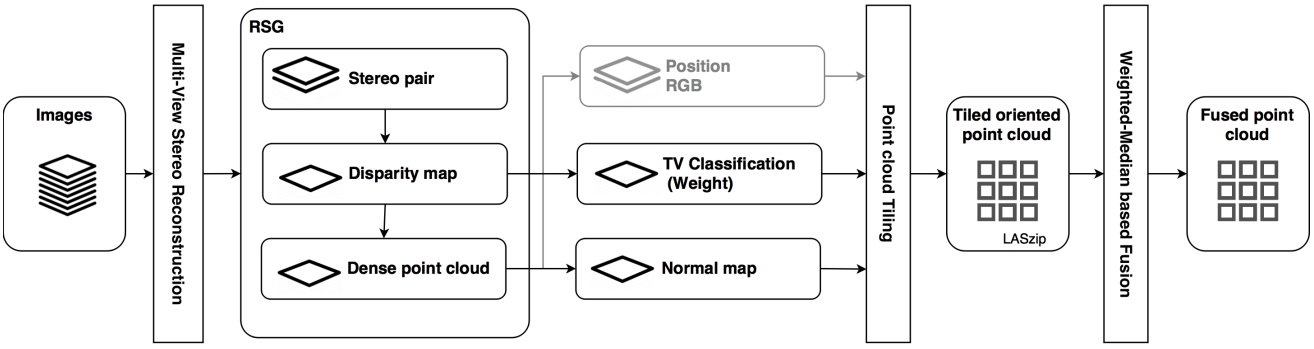

Fig. 1. Workflow of the processing pipeline for point cloud fusion.

As mentioned before, Rothermel *et al.* [15] fuse depth maps in terms of oriented 3D point cloud generation. They introduce a local median-based fusion scheme which is robust to outliers and produces surfaces comparable to the results of the Middlebury MVS. Similar to Fuhrmann and Goesele [3] points are subsampled using a multi-level octree. Favoring points with the smallest pixel footprint, an initial point set is created utilizing nearest neighbor queries optimized for cylindrical neighborhoods, points are then iteratively filtered along line of sight or surface normals. The capability of the fusion strategy for large scale city reconstruction and the straight forward manner for implementation make it particularly interesting for this work. In our work we adopt the concept of the fusion strategy using a weighted median approach favoring high quality disparities assessed by a total variation based classification.

III. METHODOLOGY

The proposed framework builds upon the Remote Sensing Software Graz (RSG)[1]. The photogrammetric processing (i.e. image registration, stereo matching) leads to different intermediate results which are utilized in the processing pipeline (see Fig. 1). Disparity maps are derived from a set of epipolar rectified images using a matching algorithm based on SGM [6]. Forward and backward matching are employed to derive two point clouds via spatial point intersection per stereo pair whose coordinates are stored in East-North-Height (ENH) raster files (i.e. a three band raster file holding the coordinates in geometry of the disparity map). The advantage of this approach is that coordinates can be accessed directly while preserving the spatial organization, i.e. the structure, of the point cloud.

In the next step, surface normals and weights are computed and stored into a compressed LAS file (i.e. a lossless compressed data format for point cloud data) [7]. Subsequently, the point clouds are assigned to tiles in order to enable a tile-wise fusion of the data. Fig. 1 depicts the complete workflow of the presented processing pipeline.

[1] `http://www.remotesensing.at/en/remote-sensing-software.html`

A. Oriented Point Cloud Generation

While in Rothermel *et al.* [15] normals are derived based on a restricted quadtree triangulation [13], we estimate surface normals in a least squares manner. A moving window operation is applied on the ENH raster files. Normals are derived by locally fitting a plane to the extracted point neighborhood. The normal estimation fails in areas with less than three reconstructed disparities. By introducing a threshold defining a minimum number of successfully reconstructed points, we are able to control the robustness of the normal calculation. In our experiments we set the pixel neighborhood to 5 pixels and used a threshold of 3 points for all datasets.

B. Disparity Quality Assessment

The quality of disparities is affected by many factors like variation of texture strength and surface slant. We assess the quality of disparities in order to derive weights for every single observed point. These weights are later used in the fusion procedure using a weighted-median approach. Kuhn *et al.* [12] introduced a TV-L^2 based classification of the disparities uncertainty. In contrast to many TV-L^1 based MVS methods, the L^2 norm takes noise and outliers into consideration which is required to measure the quality of the disparities. The TV is calculated over square windows with increasing radius m resulting in $n \in [1, 20] \subset \mathbb{N}$ discrete classes. Starting from a neighborhood containing 8 connected pixels at a radius of $m = 1$ it increases by the factor of $8m$. The discretization is achieved by introducing a regularization term τ which limits the TV to stay below a certain value. These TV classes describe the degree of the local oscillation of the disparities. The outlier probability can be obtained by learning error distributions from this classification using ground truth disparities. In our case we evaluate the quality of the disparities based on the work of Kuhn *et al.* [12] using a regularization term of $\tau = 2$.

Due to the lack of ground truth disparities, we are not able to learn error distributions directly. Therefore, we analyze the quality of the classified disparities in 3D space. Reference data from Terrestrial Laser Scanners (TLS) is used to assess the quality of the raw dense point cloud for every single TV class independently. According to Cavegn *et al.* [2], vertical Digital Surface Models (DSM) are computed for

facade patches where reference data is available. Analysing the DSM derived from the classified pointcloud and the reference data enables us to compute the weights in form of a weighting function. The weighting function is derived by calulating the standard deviation of the flatness error and fitting an exponential function in a least squares manner. The flatness error is defined as the point cloud deviations to a best fitting plane and is also an indicator for the noise of the 3D geometry [1].
Later on, we evaluate the fused pointcloud in a similiar way, to gain insight on the potential and quality of the entire fusion method. Specific information regarding the evaluation routine, selected test areas and datasets are given in Section IV.

C. Weighted-Median Based Fusion

The concept of median-based fusion originates from fusion algorithms for the generation of 2.5D DSMs. Rothermel *et al.* [15] adapted the idea by fusing point clouds in 3D space along a defined filtering direction. While for close range datasets the line of sight is suitable as filtering direction, point-wise normals are used for the fusion of aerial datasets. We adapt this fusion strategy using a weighted-median based approach.
In a first step, an initial pointset P is created from the input point cloud by storing the input point cloud in an octree data structure. The pointset P is derived by subsampling the point cloud with the centroid of the points located in a leaf node. In our work the entire fusion process was realized with the aid of the Point Cloud Library (PCL ver. 1.8.0) [16] which also provides a custom tailored octree implementation.
As a result of the disparity quality assessment every point possesses a weight representing the quality of the point. We add up the weights of all points located in the same leaf node. Thus, the weight of the initial point $p \in P$ is an indicator for the density and quality of the reconstructed scene.
Subsequently, the point cloud is fused using nearest neighbor queries optimized for cylindrical neighborhoods. For every point in the initial pointset P a set of candidate points Q, located in a cylinder with its central axis given by the initial point and its normal, is derived. Points with surface normals diverging more than 60° are discarded for further processing. After the candidate pointset Q is detected, the point p is filtered by projecting all candidate points onto the surface normal of the initial point p. Taking the weighted-median of all deviations to the point p yields the new point coordinates. Especially for noisy data further iterations can be inevitable to generate a consistent surface representation. Between every iteration, duplicate points are united to avoid redundant computations. A detailed description of the original fusion routine including the parameters and employed neighborhood queries is given in [15].
In a first iteration, Rothermel *et al.* [15] includes all points of the input point clouds for the identification of the candidate pointset Q. To speed up further iterations the filtering is restricted to the initial pointset $p \in P$ solely. In our case, we restrict the filtering of the point cloud to the initial pointset P from the beginning on. We compensate the loss of detail of the input point cloud by approximating the density of the captured 3D scene with the accumulated weight. The final surface representation is derived by discarding points with weights smaller than a defined threshold α. The influence of the threshold is analyzed in Section IV-A. In this way large and highly redundant 3D point clouds can be fused in moderate time (e.g. processing 2.5 billion points on a computer with 16 cores within a single day, resulting in a fused point cloud whose density fits the spatial resolution of the input imagery).

IV. RESULTS

In this section we discuss results obtained with the proposed fusion pipeline. The datasets used for the evaluation are provided by the ISPRS/EuroSDR project on "Benchmark on High Density Aerial Image Matching"[2] and consist of one nadir and one oblique dataset.

A. Oblique Aerial Imagery

The oblique imagery dataset was acquired over the city of Zürich with a Leica RCD30 Oblique Penta camera consisting of one nadir and four oblique 80 megapixel camera heads. While the nadir camera head is pointing downwards, directly towards the earth, the four oblique camera heads are tilted at an angle of 35 degrees, each pointing in a different cardinal direction. The entire datasets comprises 135 images, captured from 15 unique camera positions. While the nadir imagery leads to a Ground Sample Distance (GSD) (i.e. the spatial resolution) of 6 cm the GSD of the oblique views vary between 6 and 13 cm. Reference data captured with terrestrial laser scans provide accurate and reliable information for the evaluation of the datasets. The evaluation was carried out by computing DSM's of different facade patches distributed over the test area. More information on the image acquisition, benchmark and reference data can be found in [2].

Photogrammetric Processing and Pre-processing. In a first step, the image registration was carried out using the interior and exterior orientation parameters provided along with the image data. Subsequently images are matched in flight direction with an overlap of 70%, resulting in a total of 314 stereo-pairs, containing approximately 10.6 billion points. After the generation of disparity maps TV classes and normal maps are computed. As mentioned in Section IV the weighting function assigns a weight to every TV class which is then used in the fusion process.

The derived weighting function is depicted in Fig. 3 and shows that a correlation between TV classes and the geometric precision (i.e. level of noise) can be verified.

[2] http://www.ifp.uni-stuttgart.de/ISPRS-EuroSDR/ImageMatching/

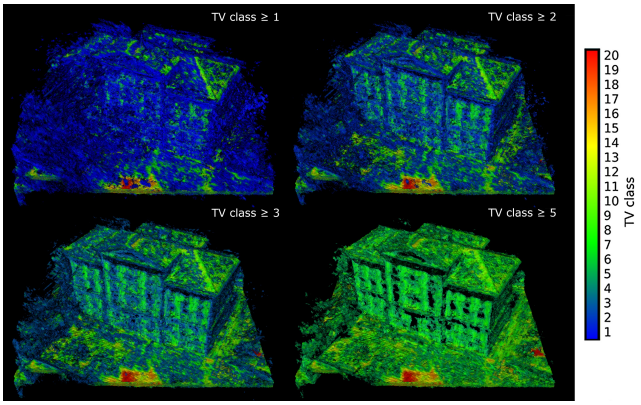

Fig. 2. Raw dense point cloud restricted to different TV classes.

While higher TV-classes show smaller standard deviations and deliver better overall accuracy, lower TV-classes are more likely to contain outliers (also cf. Fig. 2). TV classes greater than 8 are only present in flat areas facing the camera position. Since we focus on the reconstruction of vertical surfaces (i.e. facades) the information obtained by the test areas is extrapolated for all TV classes. The weighting function is derived by inverting the estimated function and defining the minimum weight with 1.0.

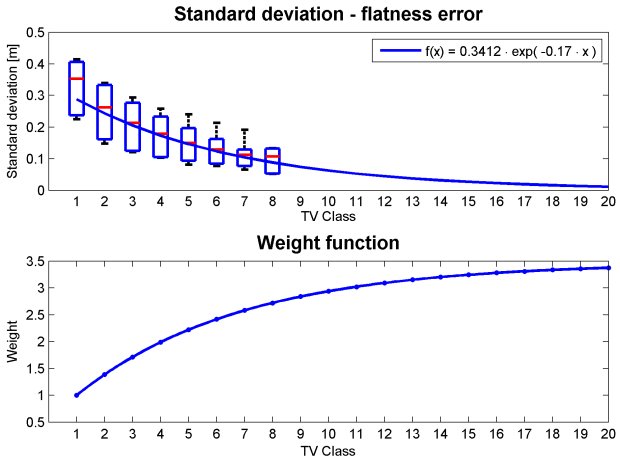

Fig. 3. Box plots representing the standard deviation of the flatness error derived from different test areas for all available TV classes *(top)*. Estimated weight function *(bottom)*.

Point Cloud Fusion. The fusion of the point cloud was carried out in three iterations with a cylinder radius of 15 cm (i.e. two times the GSD) and a height of 1.5 m. It is worth mentioning that, in some cases, during the image acquisition parts of the helicopter skids protruded into the camera angle, which leads to strong distortions in the matching procedure. The size of the octrees leaf node, which is used for the generation of the initial pointset, controls the approximate output density of the fused point cloud. Therefore, faster runtimes can be achieved producing point clouds with lower density. The resolution used for the oblique imagery is set to 10 cm, to match the GSD of the input data. Within the point cloud fusion process, the points are filtered along the surface normal and weights are accumulated. The final surface representation is derived by discarding low weights, which are more likely to contain outliers. As depicted in Fig. 4, increasing the minimum weight threshold α leads to more accurate, however less dense point clouds.

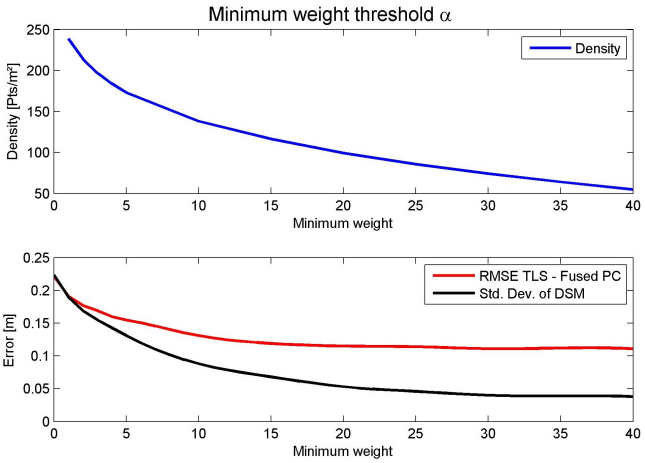

Fig. 4. Impact of rejecting low weighted points after the fusion procedure on density *(top)*, accuracy and precision *(bottom)*.

Since the fusion method produces oriented point clouds, a mesh representation can be computed using Poisson surface reconstruction [8]. The complete workflow is depicted in Fig. 5. The runtime of the fusion process can be improved by discarding low level TV classes in a pre-processing step. However, the rejection of low level TV classes causes a loss in detail in areas with bad coverage.

Evaluation. In order to measure the capability of the fusion routine different statistical measures are analyzed. The RMSE of the deviations between the reference point cloud and fused point cloud, give information about the accuracy of the 3D geometry. The standard deviation of the vertical digital surface model indicates the noise level of the point cloud, respectively the distribution of points perpendicular to the facade. As mentioned before, the density can be controlled by setting the octree resolution and by regulating the threshold for the minimum weight α. In Table I the raw point cloud is compared to the fused point cloud considering the influence of TV weights. The minimum weight threshold α is set to generate point clouds with comparable densities. Test areas include the school building located in the northern part of the mapped scene and the tower building located in the south.

TABLE I

COMPARISON OF THE FUSION ROUTINE REGARDING WEIGHTS.

	min. weight α	Density [pnts/m^2]	RMSE Fused PC-TLS [m]	Mean Fused PC-TLS [m]	Std. Dev. of DSM [m]
Raw (unfused)	-	4398.00	0.199	0.108	0.296
Fused (no weights)	20	75.15	0.122	0.067	0.052
Fused (weighted)	30	74.25	0.111	0.063	0.040
Fused (weighted pre-filter TV >1)	18	75.23	0.102	0.049	0.032

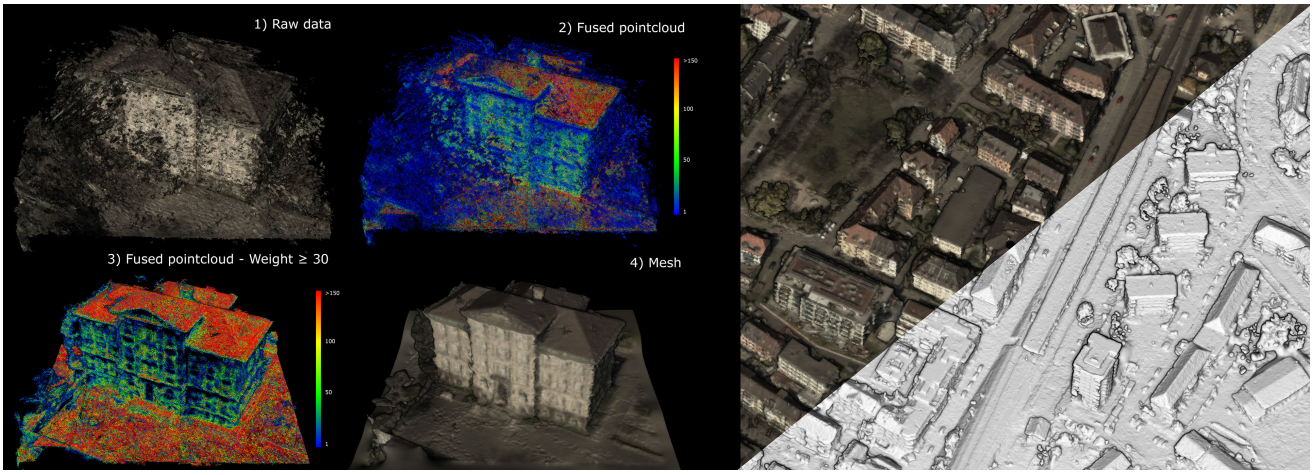

Fig. 5. Processing pipeline of the point cloud fusion: *(1)* Raw data from dense image matching (50.64 M points), *(2)* fused point cloud (1.73 M points), *(3)* discarding weights smaller than $\alpha = 30$ (0.47 M points), *(4)* mesh generation, and *(right side)* merged surface tiles.

Regarding the oblique dataset, best results can be achieved by neglecting points with TV class 1. By doing so, execution time is speed up by a factor of 2.2. Compared to the raw point cloud the fusion procedure reduces noise while improving the accuracy of the point cloud (see Fig. 6). A visual assessment shows that the fused point cloud including all TV classes and applying weights produces the best results regarding completeness and outliers (see Fig. 7). As expected, roof structures and other nadir oriented faces are reconstructed with the highest precision. Table II shows that in all cases the precision of the point cloud can be improved while decreasing redundant information.

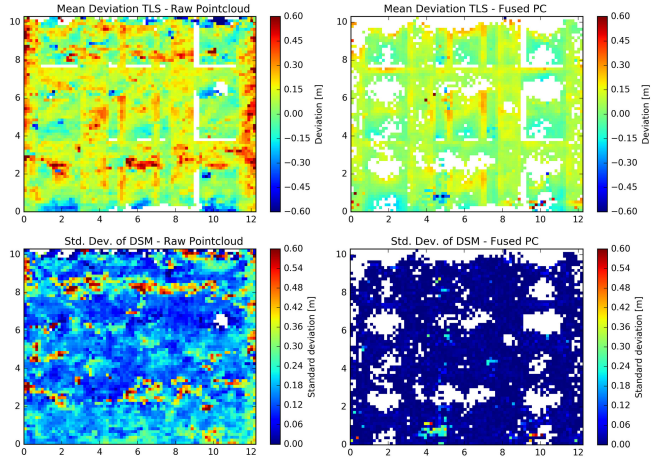

Fig. 6. Comparison of the main school facade before and after fusion procedure (cf. Fig. 5): Mean deviation between DSM derived from terrestrial laser scanner data and point cloud *(top)*, and standard deviation of the point clouds DSM representing the level of noise *(bottom)*.

Fig. 7. Taking all TV classes into account produces point clouds containing less outliers *(left)*, in contrast to point clouds restricted to TV classes > 1 *(right)*.

TABLE II

COMPARISON OF TEST AREAS BEFORE AND AFTER THE POINT CLOUD FUSION.

	Density [pnts/m^2]	RMSE Fused PC-TLS [m]	Mean Fused PC-TLS [m]	Std. Dev. of DSM [m]
Tower South (raw)	2345.9	0.378	0.051	0.538
Tower South (fused)	49.4	0.204	0.003	0.087
Tower North (raw)	1781.4	0.427	-0.222	0.447
Tower North (fused)	45.3	0.195	-0.052	0.071
Tower West (raw)	3570.8	0.350	0.237	0.499
Tower West (fused)	62.7	0.256	0.152	0.155
Roof (raw)	13864.2	0.150	-0.023	0.218
Roof (fused)	178.7	0.122	0.028	0.105

B. Nadir Aerial Imagery

The nadir image dataset covers an area of approximate 1.5 $\times$ 1.7 km^2 in the city of Munich. The dataset was acquired by a DMC II 230 megapixel aerial image camera with a spatial resolution of 10 cm and consists of 15 panchromatic images. As depicted in Fig. 8, facade information can be reconstructed by utilizing the proposed fusion routine. Due to the wide angle of the aerial camera, enough information is captured to produce 3D city models from nadir aerial imagery.

V. CONCLUSION

A novel method for fusing 3D point clouds was presented. The underlying point clouds originate from stereo matching of aerial images and were enriched by the calculation of surface normals and a classification of the disparity maps into quality classes. The proposed filtering method then fused the point cloud in direction of the surface normals and used a weighting based on the classification. Evaluation to ground truth data showed the increased quality of the fused point cloud while reducing the redundancy. Overall,

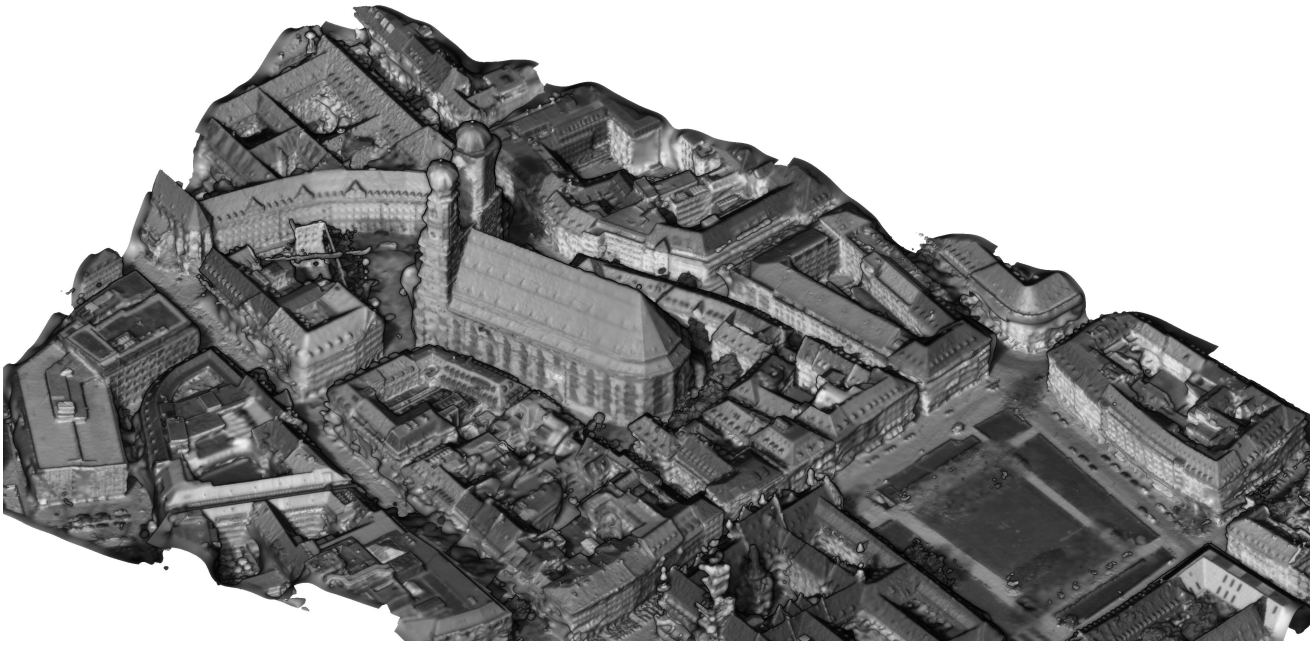

Fig. 8. Reconstructed surface from nadir aerial imagery. The depicted surface shows the Frauenkirche in Munich, located in the west part of the test area. Therefore, west-facing facades cannot be reconstructed.

this fusion concept can be easily put into state-of-the-art mapping pipelines, is able to handle large point clouds due to the tiling concept and can be applied for terrestrial, aerial or satellite based mapping application.

ACKNOWLEDGMENT

This research was partly funded by BMVIT/BMWFW under COMET programme, project nr. 836630, by Land Steiermark through SFG under project nr. 1000033937, and by the Vienna Business Agency.
The authors would like to thank Stefan Cavegn and Norbert Haala for providing the terrestrial laser scanner reference data.

REFERENCES

[1] A. H. Ahmadabadian, S. Robson, J. Boehm, M. Shortis, K. Wenzel, and D. Fritsch, "A comparison of dense matching algorithms for scaled surface reconstruction using stereo camera rigs," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 78, pp. 157 – 167, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271613000452

[2] S. Cavegn, N. Haala, S. Nebiker, M. Rothermel, and P. Tutzauer, "Benchmarking High Density Image Matching for Oblique Airborne Imagery," *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 45–52, Aug. 2014.

[3] S. Fuhrmann and M. Goesele, "Fusion of depth maps with multiple scales," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 148:1–148:8, Dec. 2011. [Online]. Available: http://doi.acm.org/10.1145/2070781.2024182

[4] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, Aug 2010.

[5] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *IEEE International Conference on Computer Vision (ICCV)*, June 2015.

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

[7] M. Isenburg, "Laszip," *Photogrammetric Engineering and Remote Sensing*, vol. 79, no. 2, pp. 209–217, 2013.

[8] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Trans. Graph.*, vol. 32, no. 3, pp. 29:1–29:13, July 2013. [Online]. Available: http://doi.acm.org/10.1145/2487228.2487237

[9] A. Kuhn, H. Huang, M. Drauschke, and H. Mayer, "Fast probabilistic fusion of 3d point clouds via occupancy grids for scene classification," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. III-3, pp. 325–332, 2016. [Online]. Available: http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/III-3/325/2016/

[10] A. Kuhn and H. Mayer, "Incremental division of very large point clouds for scalable 3d surface reconstruction," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec 2015, pp. 157–165.

[11] A. Kuhn, H. Hirschmüller, and H. Mayer, *Multi-Resolution Range Data Fusion for Multi-View Stereo Reconstruction.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 41–50. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40602-7_5

[12] A. Kuhn, H. Hirschmüller, D. Scharstein, and H. Mayer, "A TV prior for high-quality scalable multi-view stereo reconstruction," *International Journal of Computer Vision*, pp. 1–16, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11263-016-0946-x

[13] R. Pajarola, "Large scale terrain visualization using the restricted quadtree triangulation," in *Visualization*, Oct 1998, pp. 19–26.

[14] R. Perko and C. Zach, "Globally optimal robust DSM fusion," *European Journal of Remote Sensing*, vol. 49, pp. 489–511, Sept. 2016.

[15] M. Rothermel, N. Haala, and D. Fritsch, "A Median-Based Depthmap Fusion Strategy for the Generation of Oriented Points," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 115–122, June 2016.

[16] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[17] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, June 2006, pp. 519–528.

[18] B. Ummenhofer and T. Brox, "Global, dense multiscale reconstruction for a billion points," in *IEEE International Conference on Computer Vision (ICCV)*, Dec 2015. [Online]. Available: http://lmb.informatik.uni-freiburg.de//Publications/2015/UB15

[19] C. Zach, T. Pock, and H. Bischof, "A globally optimal algorithm for robust tv-l1 range image integration," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2007, pp. 1–8.