Probabilistic Range Image Integration for DSM and True-Orthophoto Generation

Markus Rumpler, Andreas Wendel, and Horst Bischof

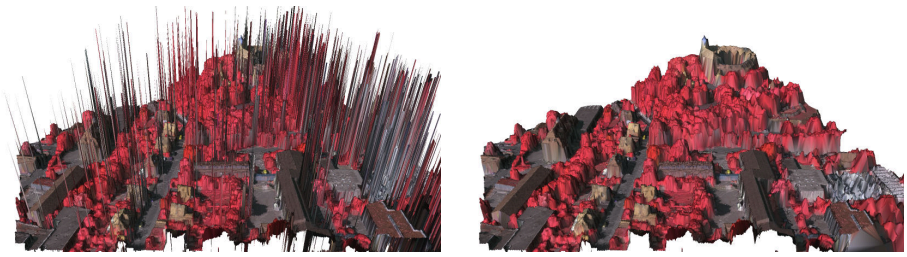
Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{rumpler,wendel,bischof}@icg.tugraz.at

Abstract. Typical photogrammetric processing pipelines for digital surface model (DSM) generation perform aerial triangulation, dense image matching and a fusion step to integrate multiple depth estimates into a consistent 2.5D surface model. The integration is strongly influenced by the quality of the individual depth estimates, which need to be handled robustly. We propose a probabilistically motivated 3D filtering scheme for range image integration. Our approach avoids a discrete voxel sampling, is memory efficient and can easily be parallelized. Neighborhood information given by a Delaunay triangulation can be exploited for photometric refinement of the fused DSMs before rendering true-orthophotos from the obtained models. We compare our range image fusion approach quantitatively on ground truth data by a comparison with standard median fusion. We show that our approach can handle a large amount of outliers very robustly and is able to produce improved DSMs and true-orthophotos in a qualitative comparison with current state-of-the-art commercial aerial image processing software.

1 Introduction

Digital surface models (DSMs) represent height information of the Earth's surface including all objects (buildings, trees, ...) on it. In geographic information systems, they form the basis for the creation of relief maps of the terrain and rectification of satellite and aerial imagery for the creation of true-orthophotos. Applications of DSMs and orthophotos range from engineering and infrastructure planning, 3D building reconstruction and city modeling to simulations for natural disaster management or wireless signal propagation, navigation and flight planning or rendering of photorealistic 3D visualizations.

A typical photogrammetric processing pipeline comprises Aerial Triangulation (AT) or Structure-from-Motion (SfM) for camera pose estimation, depth estimation by dense image matching, and range image fusion for the integration of multiple 2.5D raw range images into a single consistent DSM. Many large-scale 3D reconstruction techniques compute a final model by merging depth hypotheses from multiple, pairwise estimated range images *a posteriori* (*i.e.* one depth map from all possible stereo pairs). This is important to create a reliable fused



(a) (b)

Fig. 1. Range image integration results on the ISPRS Vaihingen test dataset [2]. (a) Fused DSM computed by median fusion, which shows considerably non-robust results. Single isolated outliers are still present in the results. (b) Our probabilistic fusion approach effectively discards outliers and produces visually clean results.

3D point cloud and a dense, watertight surface mesh. The fusion step is influenced by accumulated registration errors of the cameras and noise from false depth estimates. These erroneous depth estimates and outliers usually lead to a poor integration and consequently, need to be handled robustly (Figure 1).

Methods for range image integration in related work operate on different representations. They can be roughly categorized into mesh-based, volumetric and point-based methods.

2.5D range images can be converted directly into polygonal range meshes. Turk and Levoy [16] describe an explicit polygonal approach for merging several depth maps by zippering together adjacent meshes into a single continuous surface. More general 3D reconstruction and surface extraction methods often utilize a volumetric representation to handle arbitrary objects [3, 19, 6]. These methods extract a final surface by computing an averaged signed distance field separately on all voxels induced by the range images. Surface regularization can be integrated to allow for spatial coherence and smoothness within a global energy minimization framework, which can then be optimized using variational techniques [20] or graph cuts [18].

Most of the methods mentioned above are, due to the limited scalability of the volumetric representation, not well suited for large scale scenes with respect to memory consumption and runtime performance. Other methods [9] try to reconstruct a surface from oriented sample points, but do not perform very well in the presence of gross outliers. Merrell *et al.* [10] tries to identify conflicts and errors for depth hypotheses by fusing information of neighboring views based on visibility violations of the free-space constraint to clean the initial depth maps. Afterwards, the unprojected 3D points of cleaned depth maps adjacent to each other are merged to extract a surface mesh.

Point-based methods try to fuse information of neighboring views locally by a voting scheme. In the simplest approach, a robust fusion with respect to some isolated outliers may be performed by selecting the median depth value of all depth estimates projecting into the same cell.

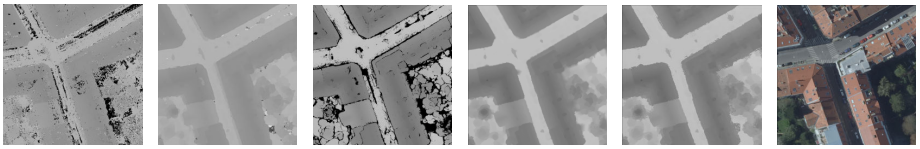
Gallup *et al.* [5] reduce the memory consumption compared to volumetric approaches by computing a height value to minimize the amount of empty space below and filled space above. The votes for empty and filled are accumulated from viewing rays of all cameras contributing to the 2D grid cell. While being parallelizable for each DSM grid cell, the final fused depth value is still determined in discrete depth steps. In the depth fusion method of Unger *et al.* [17], a number of pair-wise disparity maps are projected into a reference view pair. They fuse them by estimating a probability density function using the reprojection uncertainties and their photo-consistencies and select the most probable one from the probability distribution.

We propose a probabilistically motivated 3D filtering scheme for range image integration. Our approach works in object-space, thus, we do not need to select a reference view. Furthermore, our method has the advantage that it avoids a volumetric voxel representation and is able to select a final depth from continuous values in object-space instead of a discrete voxel sampling. For each DSM grid cell, we estimate a probability density function of the depth estimates and select the most reliable one from the probability distribution. The method is memory efficient and can easily be parallelized for each individual DSM grid cell. Furthermore, it is able to integrate point clouds from arbitrary sources into one consistent DSM. An additional photometric refinement of the DSMs based on a multi-image correlation score exploits neighborhood information from a Delaunay triangluation of the DSM grid points.

2 Range Images from Dense Image Matching

We employ a fully automatic processing pipeline that takes high resolution images as input for aerial triangulation (AT), also known as Structure from Motion (SfM), estimating the unknown camera orientations and a sparse reconstruction of the scene. For photogrammetric end-products like digital surface models or true-orthophotos, dense 3D geometry is required. In contrast to traditional approaches for dense 3D reconstructions that are based on stereo pairs [7], we employ a multi image matching method based on a plane sweep approach [1] for depth estimation.

The plane sweep approach enables an elegant way for image based multi-view reconstruction since image rectification is not required. The method allows the reconstruction of depth maps from arbitrary collections of images and implicit aggregation of multiple view's matching costs. Furthermore, the method is well suited to run on current graphics processing units (GPUs) which makes processing of large datasets consisting of many high resolution aerial images feasible in reasonable time [8]. To achieve high performance and low memory requirements, final depth values are extracted from the volume by a simple winner-takes-all (WTA) [14] strategy. For considering neighborhood information for depth extraction, fast cost volume filtering [13] or global optimization methods [8] can be applied. Figure 2(a) shows resulting range images computed by winner-takes-



(a) (b) (c) (d) (e) (f)

Fig. 2. Depth maps, fusion results and true-orthophoto from the same region. (a) Part of a range image computed with our multi-view plane sweep approach using winner-takes-all (WTA) with many outliers. (b) The same image region with a local edge preserving filter applied on the costs before WTA depth extraction. (c) Fused depth estimates with high confidence in regions where the aerial images show sufficient texture and intensity gradients, *e.g.* at corners, edges, depth discontinuities and contours. (d) A dense DSM is obtained by Delaunay triangulation in 2D, where missing parts are filled by the implicit regularization characteristic of the triangles. (e) DSM after photometric refinement; depth edges appear sharper. (f) True-orthophoto.

all and range images obtained with cost volume filtering (Figure 2(b)), which applies a local edge preserving filter on the costs before WTA depth extraction.

In this paper, we focus on fast processing of the range images and therefore use the local cost volume filtering approach. This local approach for depth extraction is, concerning runtime, 25 times faster compared to the global method [8]. Besides the longer runtime, the regularization of the global method leads to early smoothing in the beginning of the pipeline, and details, once smoothed, cannot be recovered later on. Although the depth maps from local cost volume filtering contain outliers, more details are preserved in the final DSMs. We demonstrate this effect in our experiments in Figure 7.

3 Probabilistic Range Image Integration

For every image, we compute a range image encoding depth information for every pixel. The depth information from all images needs to be integrated robustly into a single, consistent DSM with reliable height values by subsequent fusion of the available depth hypotheses.

The quality of the individual range images depends on the input data and the method used for image matching and depth extraction. Hence, the resulting depth maps may contain noise and gross outliers due to erroneous depth values. These outliers can originate from image noise, illumination changes and shadows, as well as from occlusions, repetitive structures or uniform image regions with insufficient texture, which usually lead to a poor integration and consequently, need to be handled robustly. We assume that a height value in the final DSM is reliable when most of the individual depth estimates vote for the same depth and are valid in most range images.

3.1 Orthographic Projection and Sampling

First, we define grid dimensions and sampling for the final DSM. For georeferenced datasets, we select the ground plane as the projection plane. For re-

constructions in arbitrary local coordinate frames, we estimate an appropriate rotation to transform the coordinate frame into an east-north-up (ENU) coordinate system with the z-axis pointing upwards and align the reconstruction with the xy-plane. The grid dimensions $(x_{min}, x_{max}, y_{min}, y_{max})$ are retrieved from the sparse scene reconstruction of the SfM result. The DSM grid sampling Δx and Δy in the horizontal xy-plane is chosen according to the ground sampling distance (GSD) of the input images or manually defined to meet a desired spatial resolution of the output DSM. In addition, we define a tiling of the complete DSM grid into smaller subregions (we choose tiles of size 2048×2048 pixels), which can be processed independently to cope with large scale reconstructions. All range images that overlap with a DSM tile, *i.e.* from cameras of which their viewing frustum intersect with the bounding volume of the sparse reconstruction within the tile, are used in the subsequent steps.

We unproject the range images to 3D point clouds and ortho-project the points such that the highest points from each individual range image falling into the same xy grid cell are inserted into bins. This orthographic projection geometrically undistorts the perspective range images, such that overly filled bins due to *e.g.* facades are avoided.

3.2 Robust Depth Fusion based on Point Density

We consider depth hypotheses from the range images to be samples from an underlying probability density function (Figure 3). Based on that, local maxima of the density function correspond to dense regions in feature space. We process each cell of the DSM tile independently, but use a 3×3 neighborhood of the individual cell as input to define the density function. This acts as a probabilistically motivated filter scheme in 3D. The modes of the density function then indicate more reliable depth hypotheses, and the approach is not affected by quantization artifacts. By finding these modes, we can estimate a robust depth supported by many individual range images.

Clustering of Depth Hypotheses. The probability density of the depth hypotheses for one bin is estimated by the local density of the sample points. For each data point, a local window (*i.e.* Parzen window, kernel) is defined and the data point is associated with the nearby local maximum of the probability density function within the window. For the kernel density estimation function, we use a Gaussian kernel of the form

$$\phi(x - x_i) = e^{-\frac{\|x - x_i\|^2}{2h^2}}, \tag{1}$$

where h is the bandwidth parameter, *i.e.* the radius of the Parzen window. We set $h = 10 \cdot \Delta x$ in our implementation.

Dense regions in feature space correspond to local maxima of the density function. To find the maxima of the locally estimated density, we perform gradient ascent for each data point. At each iteration, the window shifts to a denser region and converges to the modes of the density function. This is exactly the *mean shift* [4] procedure. For all data points, the mean shift vector is computed

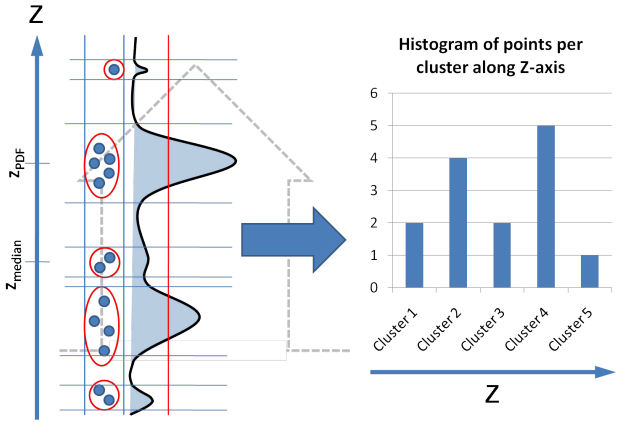


Fig. 3. Probabilistic range image fusion: Depth hypotheses are considered as samples from an underlying probability density function, where local maxima of the density function correspond to dense regions in feature space. The local maxima correspond to point clusters with higher density for which we build a histogram. Median fusion of all depth hypotheses would lead to a wrong height value whereas the probabilistic approach is able to select a height value correctly.

in every iteration and the density estimation window moved until convergence. The mean shift vector for a data point x is computed as

$$m(x) = \frac{\sum_{i=1}^{n} x_i \phi(x - x_i)}{\sum_{i=1}^{n} \phi(x - x_i)}. \tag{2}$$

All data points that converge to the same mode are associated to belong to the same cluster. The mode corresponding to a cluster defines the 'center of mass' for the data points.

Mode Selection. We build the histogram of sample points per cluster (*i.e.* depth hypotheses) and sort the modes by their height value in ascending z-direction (see Figure 3). We accept only clusters with a minimal support of two points. The fused height value for a cell is then selected based on the maximum peak in the points histogram. If more than one identical peaks are present in the histogram, we select the spatially highest mode such that the empty space criterion is not violated.

3.3 Mesh Generation

Cells for which we cannot find at least one sufficiently dense cluster with minimum support are discarded. We consider them as unreliable and keep the corresponding DSM cell empty. This results in a semi-dense DSM height map as depicted in Figure 2(c), where only sample points of high confidence and support from many range images are kept. Those fused DSM points are then triangulated in 2D by a Delaunay triangulation, extruded to their estimated height and converted into a polygonal mesh. The mesh representation offers several favorable

properties including data reduction and automatic elimination of holes caused by the discarded depth estimates with low support. The missing parts in the DSM are filled by the implicit regularization characteristic of the triangle between the high-confidence fused sample points through linear interpolation.

Such an interpolation is possible because high confidence points are present in regions where the aerial images show sufficient texture and intensity gradients, which is especially the case at corners and edges, caused by depth discontinuities and object contours; *i.e.* where the NCC matching finds distinctive matches. In contrast, points in homogenous, typically planar regions are suppressed. This effect has the additional advantage of simplifying the 3D models such that planar regions are not necessarily densely sampled but simplified, and thus achieving a data reduction. The polygonal mesh representation is ideally suited as it allows a compact, memory efficient representation of irregular structure and varying complexity. It can be easily distributed and directly visualized in 3D (see Figure 4(a)).

4 DSM Refinement and True-Orthophoto Generation

Finally, we describe how neighborhood information given by the Delaunay Triangluation can be exploited for photometric refinement of the fused DSMs, and how to render true-orthophotos from the obtained meshes.

4.1 Photometrically Constrained Refinement

For each DSM grid cell where no reliable depth estimate was found, mesh triangles may span over large gaps of uncertain regions (Figure 2(d)). We employ ray casting for each DSM grid cell and intersect the ray in z-direction with the intermediate triangular mesh. If the intersected triangle spans over large depth discontinuities (*i.e.* if the triangle vertices have significantly varying heights and the deviation between triangle normal and ray is larger than 30°), we allow the height of the triangle intersection point to jump to one of its neighbor's height values. To decide which height value is best, we estimate a multi-image correlation score [11] for the backprojected intersection point at all three depths of its neighboring triangle vertices and allow the interpolated depth of the intersection point to either jump to the height of the most reliable neighbor or stay at the same height, depending on where the best photo-consistency score is found. Hence, we let neighboring height values (defined by the triangulation) propagate into areas with large uncertainty. After photometric refinement, the depth edge appears sharper (Figure 2(e)).

4.2 Rendering True-Orthophotos

True-orthophotos are a common end-product obtained from digital surface models where the aerial images are rectified from a perspective to an orthographic projection using an underlying DSM (Figure 4).

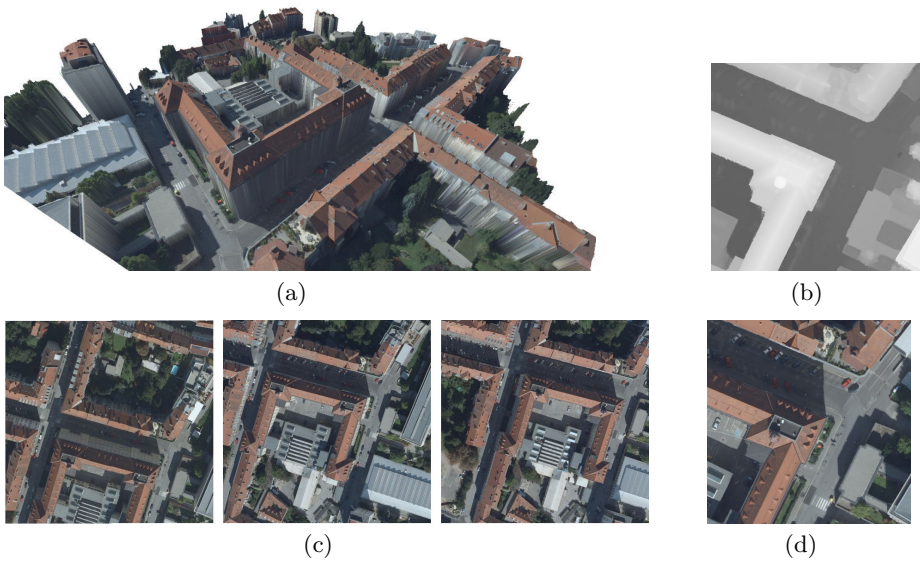
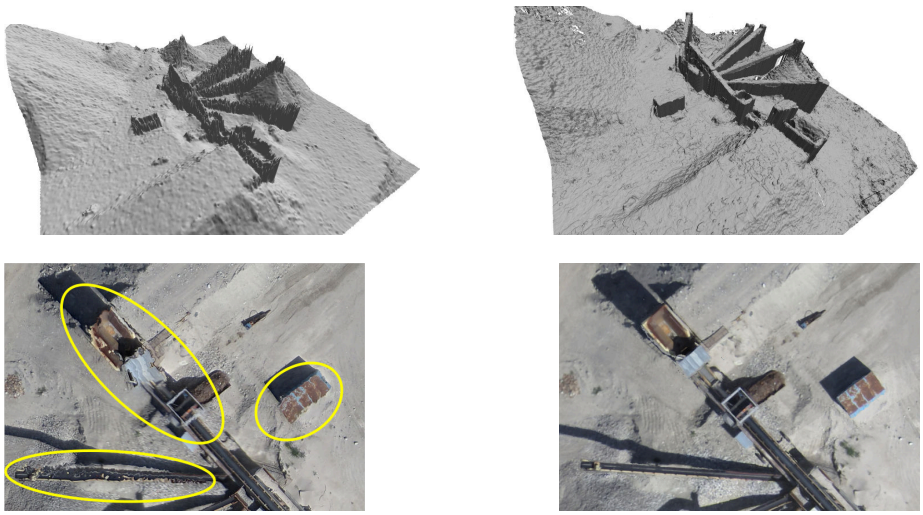


Fig. 4. (a) 3D visualization of the resulting DSM as a triangular mesh and textured with the true-orthophoto. (b) DSM represented as a height map. (c) Perspective views of the building from the original aerial images. Note that building facades are visible. (d) True-orthophoto where no facades are visible, generated from the aerial images and the underlying DSM heights.

Again, orthographic ray casting is used to sample the mesh. For each DSM grid cell, the intersection point gets back-projected into all images where it is visible. For each DSM point we obtain a set of color values projecting to that point. In many cases, the color values are in RGB space and simple averaging would lead to inappropriate color shifts in the resulting true-orthophoto. Furthermore, illumination changes in images from different flight stripes, taken at different times of the day may also introduce undesired color artifacts. We solve this issue by transforming the color values to HSI color space, which splits the color components into hue, saturation and intensity such that the color value is independent from saturation and brightness. We finally assign the color triplet where the intensity corresponds to the median intensity of all color samples.

5 Results and Discussion

We compare our range image fusion and true-orthophoto generation approach qualitatively with results from a state-of-the-art commercial aerial image processing software (Pix4D [12]) on image test sets taken from different unmanned aerial vehicles (UAVs), and on a large scale dataset with 21 RGB images captured with an UltraCam X with a resolution of 14,430 x 9,420 pixels. We also perform a quantitative evaluation on the ISPRS Vaihingen test dataset [2] with 20 infrared-red-green (IR-R-G) DMC images comprising a resolution of 7,680 x 13,824 pixels for which a DSM with 25 cm resolution from airborne laser scanning (ALS) data



(a) (b)

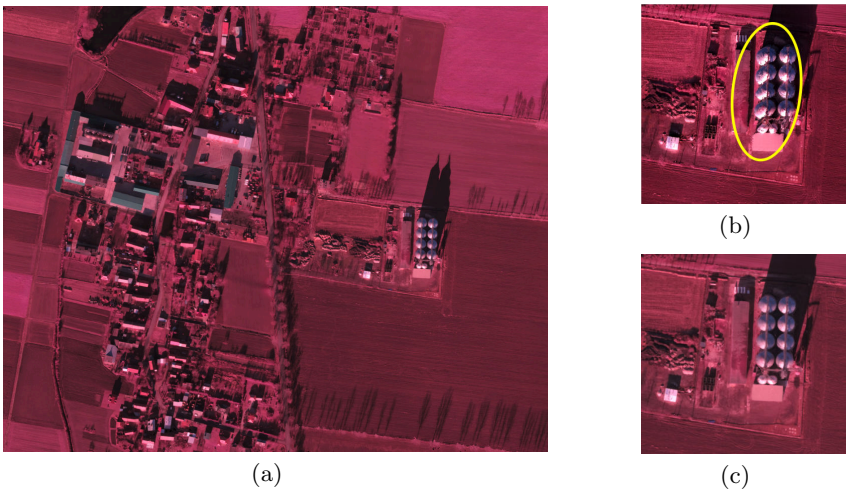
Fig. 5. Qualitative comparison between results from the commercial software Pix4D and our approach. Datasets were taken from [12]. (a) 3D view and true-orthophoto of a part of the 'gravel pit' dataset, which shows disturbing spikes and significant noise on the band conveyor. Thus, erroneous distortions in the orthophoto (highlighted in the image) occur. (b) In contrast, our results are well fused and show sharp object contours.

is provided. Both high resolution datasets provide an image overlap of 60 - 80% and a ground sampling distance (GSD) of 8 cm/pixel.

Figure 4 shows results obtained with our probabilistic range image integration approach from winner-takes-all depth maps. The raw WTA range maps contain lots of outliers, but our robust fusion result visualized as a 3D mesh shows virtually no severe outliers. The final DSM generated from the UltraCam X dataset shows an area of approximately 3.46 km^2 and has a resolution of 20,699 x 26,096 pixels with 8 cm GSD. The GPU-based calculation of all 21 depth maps with the local method required 9 hours, the fusion took 15 hours using four CPU cores. In comparison, the global method requires more than 12 days.

A qualitative comparison between DSMs and true-orthophotos generated with the commercial software Pix4D [12] and our approach is given in Figures 5 and 6. The quality of the true-orthophotos directly depends on the correctness of the underlying DSM. Wrong depths cause distortions in the generated orthoimages as can be seen in the examples. Our method for true-orthophoto rendering has the advantage that it operates fully automated and avoids (often manual) seamline selection and ghosting artifacts in contrast to orthophotos generated from image mosaics as applied in many software tools.

For a quantitative evaluation, we compared the results of our probability based clustering fusion approach to a median fusion of the range image depth



(a) (b) (c)

Fig. 6. (a) True-Orthophoto showing the complete area of the 'Near Infrared Agriculture' dataset from [12] calculated with our approach. (b) Crop from the Pix4D orthophoto shows erroneous distortions due to errors in the underlying DSM (highlighted in the image). In contrast, our result (c) shows no distortions.

hypotheses. The range images were computed using cost volume filtering [13] before performing winner-takes-all. The chosen evaluation strategy is similar to [15]. Although, we achieve only 5% better results than median fusion, a direct visual comparison between median fusion and our proposed method clearly shows a significant difference. A small fraction of outliers which do not have a big impact in the quantitative analysis can have a drastic impact on the qualitative appearance, as shown in Figure 1.

There is a trade-off between putting large effort in generating clean, outlier-free range maps using global optimization methods and fast depth estimation using local methods which permit a certain amount of outliers in the data but preserve detail. Once the input data is smoothed it may be fused using a simple method like median fusion, but with the drawback of not being able to recover fine detail (see Figure 7). In contrast, our method can deal with a substantial amount of outliers in the range data as long as one distinct dense point cluster can be estimated, and it is considerably faster.

6 Conclusion

We proposed a probabilistic method for range image integration and true-orthophoto generation that considers depth hypotheses as samples from an underlying probability density function. The modes of the density function indicate more reliable, robust depth hypotheses supported by many individual range images. Neighborhood information given by a Delaunay triangulation can

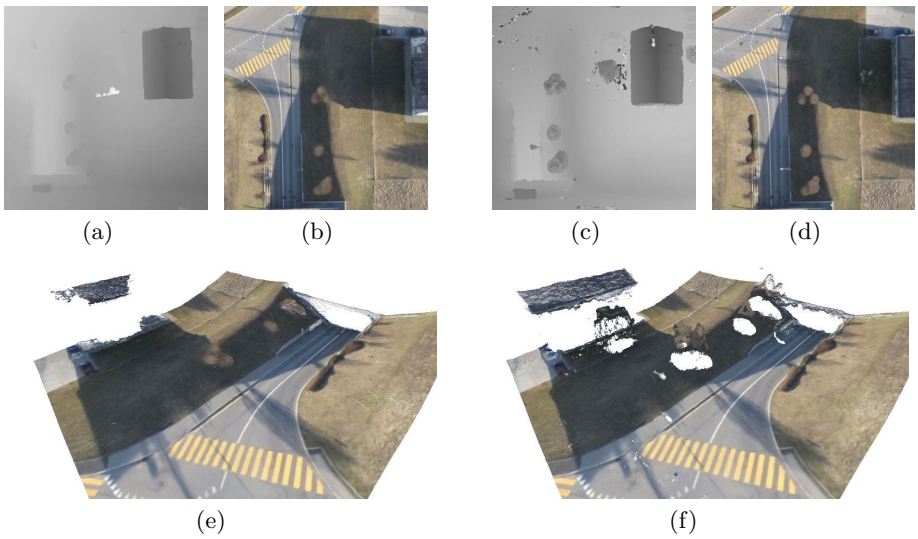


Fig. 7. Comparison between different methods for depth estimation and fusion on the 'Showcase 1' datasets from [12]. (a) Depth map calculated using the method of [8] with global optimization, which produces nearly outlier free but over-smoothed depth maps. Details like the trees and street lights are lost in the fused DSM using simple and fast median fusion, visualized in (e). The depth map in (c) has been estimated using fast cost volume filtering [13]. It contains a few outliers, but preserves details. Our probabilistic range image fusion method is able to handle large amounts of outliers and still produces visually competitive DSM quality (f). True-orthophotos generated from globally optimized depth maps (b) and cost volume filtering (d) look very similar, but again the details are better preserved by our method (e.g. the roof of the house).

be exploited for photometric refinement of the fused DSMs before rendering true-orthophotos from the obtained meshes.

Our approach has the advantage that it avoids a volumetric voxel representation and is able to select a final depth from continuous values instead of a discrete voxel sampling. The method is memory efficient and can easily be parallelized for each individual DSM grid cell. Furthermore, our approach enables to integrate arbitrary point sets into one single 2.5D representation. It considers the unprojected 3D points only, hence, it is not restricted to range image integration but is able to fuse *e.g.* point clouds provided by LiDAR into a consistent DSM. Even though the simplicity of our method, we showed that our approach is very robust with respect to outliers in the depth estimates and is able to deal with a substantial amount of outliers in the range data to produce good results.

Acknowledgements. This work has been supported by the Austrian Research Promotion Agency (FFG) FIT-IT project HOLISTIC (830044). The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [2]: http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html (in German).

References

1. Collins, R.T.: A Space-Sweep Approach to True Multi-Image Matching. CVPR (1996)
2. Cramer, M.: The DGPF-Test on Digital Airborne Camera Evaluation – Overview and Test Design. Photogrammetrie - Fernerkundung Geoinformation (2010)
3. Curless, B., Levoy, M.: A Volumetric Method for Building Complex Models from Range Images. SIGGRAPH (1996)
4. Fukunaga, K., Hostetler, L.D.: The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. IEEE Transactions on Information Theory 21 (1975)
5. Gallup, D., Frahm, J.-M., Pollefeys, M.: A Heightmap Model for Efficient 3D Reconstruction from Street-Level Video. 3DPVT (2010)
6. Hilton, A., Stoddart, A.J., Illingworth, J., Windeatt, T.: Reliable Surface Reconstruction from Multiple Range Images. ECCV (1996)
7. Hirschmueller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. CVPR (2005)
8. Irschara, A., Rumpler, M., Meixner, P., Pock, T., Bischof, H.: Efficient and Globally Optimal Multi View Dense Matching for Aerial Images. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2012)
9. Kazhdan, M., Bolitho, M., and Hoppe, H.: Poisson Surface Reconstruction. Symposium on Geometry Processing (2006)
10. Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.-M., Yang, R., Nistér, D., Pollefeys, M.: Real-time Visibility-based Fusion of Depth Maps. CVPR (2007)
11. Paparoditis, N., Thom, C., Jibrini, H.: Surface Reconstruction in Urban Areas from Multiple Views of Aerial Digital Frame Cameras. International Archives of Photogrammetry and Remote Sensing (2000)
12. Pix4D Aerial Image Processing Software, `http://www.pix4d.com` (January 2013)
13. Rhemann, C., Hosni, A., Bleyer, M., Rother, V., Gelautz, M.: Fast Cost-Volume Filtering for Visual Correspondence and Beyond. CVPR (2011)
14. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. Int. Journal of Computer Vision 47 (2002)
15. Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery. CVPR (2008)
16. Turk, G., Levoy, M.: Zippered Polygon Meshes from Range Images. SIGGRAPH (1994)
17. Unger, C., Wahl E., Sturm P., Ilić, S.: Probabilistic Disparity Fusion for Real-Time Motion-Stereo. Technical Report, TUM (2010)
18. Vogiatzis, G., Torr, P., Cipolla, R.: Multi-View Stereo via Volumetric Graph-Cuts. CVPR (2005)
19. Wheeler, M., Sato, Y., Ikeuchi, K.: Consensus Surfaces for Modeling 3D Objects from Multiple Range Images. ICCV (1998)
20. Zach, C., Pock, T., Bischof, H.: A Globally Optimal Algorithm for Robust TV-L^1 Range Image Integration. ICCV (2007)