FAST PROBABILISTIC FUSION OF 3D POINT CLOUDS VIA OCCUPANCY GRIDS FOR SCENE CLASSIFICATION

Andreas Kuhn[a,b], Hai Huang[a], Martin Drauschke[a,b], Helmut Mayer[a]

[a] Bundeswehr University Munich - (andreas.kuhn, hai.huang, helmut.mayer)@unibw.de
[b] German Aerospace Center - martin.drauschke@dlr.de

Commission III, WG III/4

KEY WORDS: Scene Classification, Point Cloud Fusion, Multi-View Stereo

ABSTRACT:

High resolution consumer cameras on Unmanned Aerial Vehicles (UAVs) allow for cheap acquisition of highly detailed images, e.g., of urban regions. Via image registration by means of Structure from Motion (SfM) and Multi View Stereo (MVS) the automatic generation of huge amounts of 3D points with a relative accuracy in the centimeter range is possible. Applications such as semantic classification have a need for accurate 3D point clouds, but do not benefit from an extremely high resolution/density. In this paper, we, therefore, propose a fast fusion of high resolution 3D point clouds based on occupancy grids. The result is used for semantic classification. In contrast to state-of-the-art classification methods, we accept a certain percentage of outliers, arguing that they can be considered in the classification process when a per point belief is determined in the fusion process. To this end, we employ an octree-based fusion which allows for the derivation of outlier probabilities. The probabilities give a belief for every 3D point, which is essential for the semantic classification to consider measurement noise. For an example point cloud with half a billion 3D points (cf. Figure 1), we show that our method can reduce runtime as well as improve classification accuracy and offers high scalability for large datasets.

1. INTRODUCTION

Scene classification is important for a wide range of applications and an open field in research concerning runtime, scalability and accuracy. Accurate 3D point clouds are essential for robust scene classification for state-of-the-art methods. It has been shown that 3D point clouds from laser sensors are suitable for this task (Schmidt et al., 2014). Unfortunately their acquisition is expensive, laser sensors are relatively heavy and have a high energy consumption.

The recent progress in image-based 3D reconstruction by Multi-View Stereo (MVS) methods allows for the generation of 3D point clouds from images, also in large numbers. High resolution consumer cameras on Unmanned Areal Vehicles (UAVs) offer a cheap acquisition of images from novel viewpoints. Unfortunately, the generation of accurate 3D point clouds requires a high computational effort.

Particularly, stereo methods like Semi-Global Matching (SGM) (Hirschmüller, 2008) can generate disparity maps also from very high resolution images. Recently, the fusion of large sets of disparity maps (billion 3D points) to accurate 3D point clouds has been demonstrated (Kuhn et al., 2013, Fuhrmann and Goesele, 2011, Fuhrmann and Goesele, 2014). Unfortunately, the processing of such large point clouds on single PCs can take a couple of days (Ummenhofer and Brox, 2015). For practical applications such as scene classification, there is basically no need for a computationally complex fusion to obtain accurate 3D point clouds. We, thus, show that a fast fusion via occupancy grids essentially speeds up the runtime and offers similar quality for semantic classification when considering probabilities.

Scene classification is an essential and intensively studied topic in photogrammetry, remote sensing and geospatial information science. Many approaches have been reported over the last decades. Sophisticated classification algorithms, e.g., support vector machines (SVM) and random forests (RF), data modeling methods, e.g., hierarchical models, and graphical models such as conditional random fields (CRF), are well studied. Overviews are given in (Schindler, 2012) and (Vosselman, 2013). (Guo et al., 2011) present an urban scene classification on airborne LiDAR and multispectral imagery studying the relevance of different features of multi-source data. An RF classifier is employed for feature evaluation. (Niemeyer et al., 2013) proposes a contextual classification of airborne LiDAR point clouds. An RF classifier is integrated into a CRF model and multi-scale features are employed.

Recent work includes (Schmidt et al., 2014), in which full waveform LiDAR is used to classify a mixed area of land and water body. Again, a framework combining RF and CRF is employed for classification and feature analysis. (Hoberg et al., 2015) presents a multi-scale classification of satellite imagery based also on a CRF model and extends the latter to multi-temporal classification. Concerning the use of more detailed 3D geometry, (Zhang et al., 2014) presents roof type classification based on aerial LiDAR point clouds.

In this paper we present a robust and efficient analytical pipeline for automatic urban scene classification based on point clouds from disparity maps, which is adapted to utilize the additional probability information for the points to improve the results.

The paper is organized as follows: In Section 2 we describe a pipeline for the fast generation of high resolution 3D point clouds. The fusion of point clouds and the derivation of per-point probabilities are given in Section 3. Section 4 examines the use of point cloud probabilities for urban scene classification. Experiments on a large dataset (see Figure 1) are presented in Section 5. Finally Section 6 gives conclusions and an outlook to future work.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-III-3-325-2016

Figure 1. Village dataset: The four images on the left each show one of 296 36-megapixel input images acquired from an Unmanned Aerial Vehicle (UAV). In the middle the noisy point cloud derived from disparity maps accumulating half a billion 3D points is given. On the right the classification results for buildings (red), ground (gray), grass (blue) and trees (green) are presented. In spite of the huge amount of input data, our method allows for the classification within a couple of hours on a single PC.

2. GENERATION OF 3D POINT CLOUDS

In this paper, we focus on the fast generation of 3D point clouds from image sets which are suitable for semantic scene classification. We demonstrate that semantic classification of buildings, vegetation and ground for a complete village, captured by hundreds of high-resolution images leading to half a billion 3D points, is possible on a single PC within a couple of hours. Hence, especially the runtime of the processing pipeline, e.g., for generation of the point cloud is important.

The first step in a dense point cloud generation pipeline is image registration which can be done fast even for thousands of wide baseline high-resolution images (Mayer, 2015). The fast processing is possible as only a fraction of the image information is needed for the estimation of camera poses. Additionally, Graphics Processing Unit (GPU) implementations can speed up the processing (Wu et al., 2011, Wu, 2013).

The next step is MVS. Disparity maps from pairs of the entire set can be generated in parallel using multi-core systems. Nonetheless, this task can be of high computational complexity. Especially SGM (Hirschmüller, 2008) has been found to successfully compute high-resolution disparity images in reasonable time still retaining small details (Hirschmüller and Scharstein, 2009). Furthermore, for SGM, publicly available (Rothermel et al., 2012), fast GPU (Ernst and Hirschmüller, 2008) and Field Programmable Gate Array (FPGA) (Hirschmüller, 2011) implementations exist. The image set from Figure 1 with 296 36-megapixel images can be processed at quarter resolution in only two and a half hours on a single PC with the FPGA. An example disparity map of an image of this set is shown in Figure 2.

Figure 2. Example image of the urban dataset (Figure 1) and the corresponding disparity map from SGM considering twelve overlapping images showing partly the same scene. The white areas represent areas filtered by consistency checks in SGM.

The final step for the generation of accurate point clouds is disparity map fusion. Even though recently scalable fusion methods have been presented (Fuhrmann and Goesele, 2011, Kuhn et al., 2013, Fuhrmann and Goesele, 2014, Kuhn et al., 2014, Ummenhofer and Brox, 2015) they are still not able to process large amounts of data, e.g., billion of 3D points, within one day on a single PC (Ummenhofer and Brox, 2015). To overcome the problem of costly MVS-based fusion of 3D point clouds, we leverage occupancy grids (Moravec and Elfes, 1985) for the fusion arguing that the redundancy and the high image resolution of specific datasets are highly useful for applications like image classification. Therefore, the fusion of 3D point clouds from disparity maps including the derivation of probabilities and their use for scene classification are the main focus of this paper.

3. FUSION OF 3D POINT CLOUDS

For an efficient scene classification from 3D point clouds it is essential to get rid of redundancy. Additionally, the 3D point cloud from disparity maps consists of noise as well as of outliers. Our goal for point cloud fusion is the applicability to semantic scene classification, because the use of high resolution images for MVS leads to point densities from which the classification does not automatically benefit. In this section we show, how point clouds from individual images can be merged very fast via octree-based occupancy grids at a resolution suitable for classification.

To this end, first the framework of occupancy grids is described in Section 3.1. This is followed by a description of an octree-based fusion of 3D points from a single disparity map and the fusion of multiple disparity maps in Section 3.2. For the latter, the occupancy grids are used to fuse the complete set of disparity maps and for the derivation of point-wise probabilities suitable for scene classification.

3.1 Occupancy Grids

Occupancy grids are especially important for real time applications and, hence, popular in the robotics community. They were introduced by Moravec and Elfes (Moravec and Elfes, 1985) and consist of a regular decomposition of an environment into a grid. Within this representation a probability is derived for individual grid cells that a cell is occupied depending on the number of measurements assigned to the cell. This can be useful for the fusion of redundant and noisy measurements and the classification of outliers, e.g., for disparity map fusion.

Redundant measurements assigned to the same cell are merged by means of probability theory. More precisely, a Binary Bayes Filter (BBF) is used for the derivation of the probability of a cell to be occupied. Initially, an inlier probability p is defined for a measurement. The measurement, e.g., a 3D point derived from

disparity and camera calibration, is transformed into the grid depending on its position. To the corresponding cell the probability p is assigned which represents the probability of the voxel to be occupied.

When multiple measurements are assigned to one voxel cell, e.g. redundant 3D points from multiple disparity maps, BBF allows for the fusion of the probabilities. To this end, the so called logprob l is defined as $l = \ln(\frac{p}{1-p})$. The fusion is conducted incrementally assuming uncorrelated data. Initially l can be set zero corresponding to an occupation probability of 0.5. The logprob at time t is defined as:

$$l_t = l_{t-1} + \ln(\frac{p}{1-p}) \, . \tag{1}$$

The incremental formulation is a crucial advantage when fusing larger sets of disparity maps, as never more than one input point has to be considered. The overall logprob for n measurements (input points) in one cell can be formulated as:

$$l = \sum_{i=1}^{n} \ln(\frac{p_i}{1-p_i}) \, , \tag{2}$$

and rises continuously with the number of input points. In our case a constant inlier probability p derived from the disparity maps is assigned to all 3D points. Hence, $\frac{p}{1-p}$ is constant and it is sufficient to calculate it only once. After fusion of the n measurements, the logprob can be transformed back to a final inlier probability as follows:

$$p = 1 - \frac{1}{1+e^l} \, . \tag{3}$$

Figure 3 demonstrates the fusion of one to three measurements considering Equations 1 to 3.

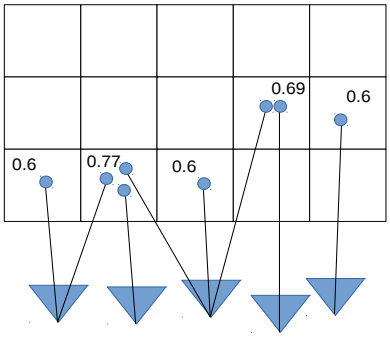

Figure 3. Assuming an inlier probability of 60% for individual points, overall inlier probabilities of 69% (two points) and 77% (three points) follow considering BBF.

For a detailed description of Occupancy Grids and the BBF see, e.g., the textbook of Thrun (Thrun et al., 2005).

3.2 3D Fusion of Multiple Disparity Maps

By considering the relative camera poses and the calibration of the camera(s), a 3D point cloud can be derived from disparity maps. Hence, the input of our proposed method is a set of point clouds corresponding to the individual disparity maps of the image set. For several applications it is useful to get rid of the high point density inherent in the disparity maps. This could be done by extraction of disparity maps from downscaled images. Because of varying distances to the scene and the resulting irregular loss in quality, we present a reduction in 3D space. As it allows parallel processing, we initially decompose the dense 3D point clouds for all disparity maps separately, to reduce the amount of data. This is our first step towards a fusion in 3D space via octrees, where the space is decomposed according to a given voxel size. For georeferenced data the voxel size can be defined by the application and the necessary accuracy, e.g. 20 cm accuracy for scene classification. Figure 4 shows an input point cloud from the disparity map shown in Figure 2 and the decomposed point cloud.

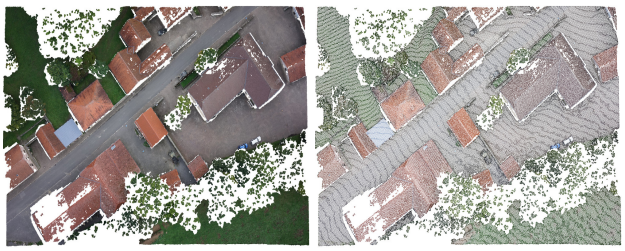

Figure 4. The left images shows the dense point cloud from the disparity map shown in Figure 2. After decomposition in 3D the density is adapted to the application of scene classification.

Octrees are suitable for fast decomposition as they offer logarithmic access time. In our method, the octree root represents the entire input point cloud. For it, a bounding volume has to be defined, which in our case is the bounding box of the input point cloud. Hence, as a first step the bounding box is calculated via the minimum and maximum x- y- and z-values of the 3D point coordinates from the input point cloud. After defining the root node, all points are traced through the octree down to the level of the decomposition size. If multiple measurements are assigned to a single voxel of the octree, the 3D coordinates are averaged. We do not use a probabilistic fusion via occupancy grids at this point, because the 3D point positions from a single disparity map are highly correlated due to the regularization terms in SGM. Hence, the geometric probabilistic fusion is only conducted on point clouds from multiple disparity maps.

Additionally to the geometric position, the color of the 3D point is essential for scene classification. To determine it, we combine multiple RGB measurements from the images in the fusion process by means of the median. Particularly, for all of three color channels the median of all measurements from one image in one cell is calculated separately. The median allows, in contrast to the mean, sharper borders between different object classes and, hence, is suitable for scene classification.

For fusion of the 3D point clouds and derivation of a pointwise probability, especially the merging of the individual (decomposed) point clouds is of interest as we assume them to be uncorrelated. To this end, we transform the reduced point clouds derived from the disparity maps into occupancy grids. For fast access times, again, octrees are used whose root size can be easily derived from the set of bounding boxes of the individual disparity maps. As in the decomposition, 3D coordinates in individual voxels are averaged while the RGB color is fused via median. Additionally, a probability is derived depending on the number of measurements.

The incremental definition of the logprob fusion in the BBF (cf. equation 1) allows for a sequential processing of the set of input point clouds. This is an important benefit, as never more than

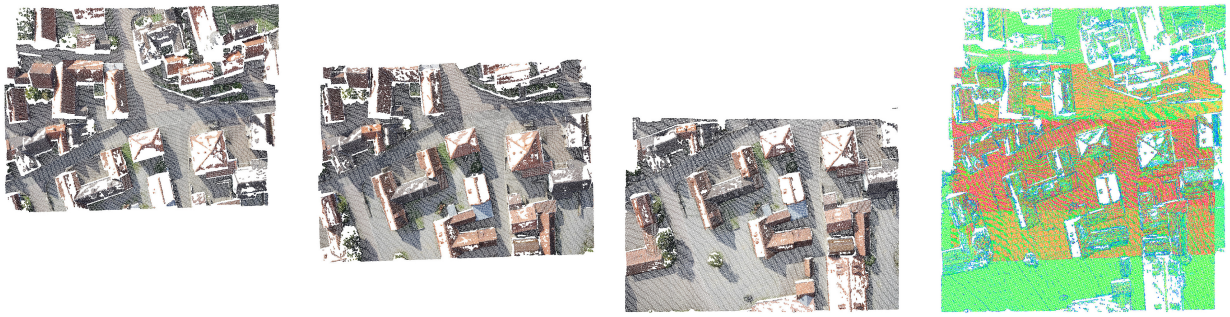

Figure 5. The left three images show the reduced point clouds derived from three disparity maps. On the right the coded point cloud representing inlier probabilities derived by means of an occupancy grid is given. The overlapping region in the center has higher probabilities as it has votes from three point clouds. Only the red points are contained in all of the three input point clouds and, hence, have the highest probability.

one input point has to be processed at a time which guarantees high scalability even for large datasets. For all point clouds the 3D points are traced down to the given octree size, which equals the size in the individual disparity map decompositions. If the assigned voxel is not occupied, the defined inlier probability p is assigned to the voxel in its logprob representation l. In case the voxel is occupied, the logprobs are merged by the incremental sum (see Equation 1). After the octree has been built from the entire set of point clouds, the final point cloud with inlier probabilities, derived from the logprobs by Equation 3 is used for scene classification.

Figure 5 demonstrates the probabilistic derivation by an example consisting of three input point clouds. The probability of 3D points in the voxels rises continuously with the number of measurements (see Figure 4).

In summary, the fusion allows for a probabilistic decomposition of redundant data for multiple disparity maps. Additionally, to the benefit of data reduction, the derived point-wise probabilities are an essential prerequisite for a derivation of a stable scene classification.

Figure 6. Comparison of the rasterized data based on conventional (left) and probabilistic fusion (right): Point clouds presented in color (top) and as elevation (bottom).

4. SEMANTIC CLASSIFICATION

The reconstructed and fused point cloud is the basis for scene classification. The approach presented in (Huang and Mayer, 2015) is extended to utilize the additional probabilities assigned to the data points. The proposed classification works on rasterized (with x-y the basis and z the elevation) and, therefore, reduced point clouds. The quality of the points indicated by the given probabilities is employed to preserve the more meaningful data points. Figure 6 compares the rasterized data derived from non-probabilistic (left) and probabilistic (right, points with the best probabilities) fusion.

4.1 Patch-wise Scheme

Patch-wise classification is inspired and employs the image segmentation with superpixels (Ren and Malik, 2003, Felzenszwalb and Huttenlocher, 2004, Moore et al., 2008). The latter is widely used in image processing, but has not been adapted and employed for 2.5D or 3D data. The main reason is that the superpixel segmentation does not consider 3D geometry. It, thus, often results in a false segmentation, i.e., superpixels that cross objects, which directly leads to errors in the final result, which basically cannot be corrected by any post-processing. To tackle this, a reliable oversegmentation with full 3D parsing of the geometry is required. We follow the segmentation idea described in (Huang et al., 2014) to group the data points. As presented in Figure 7, an extended super-pixel segmentation for multiple channels including RGB, DSM and local normal directions is performed to oversegment the data into 3D "patches". Data points inside one patch are homogeneous concerning color and 3D geometry, which implies they belong to the same object.

The advantage of the patch-wise scheme lies in its efficiency: Both the feature calculation and the classification only need to be conducted once and can be applied for all members of the same patch. Please note, however, the key that makes the fast scheme also achieves an acceptable classification accuracy is an appropriate oversegmentation. The improved color and elevation values, as shown in Figure 7, lead to a better segmentation with clearer boundaries and, therefore, ensure the feasibility of the patch-wise scheme.

4.2 Relative Features

"Relative" features instead of absolute ones lead to a more stable classification. Relative heights of buildings and trees in relation to the ground can be derived based on an estimated DTM (digital terrain model). The classification, however, still suffers from (1) the heterogeneous appearance of the objects in the same class,

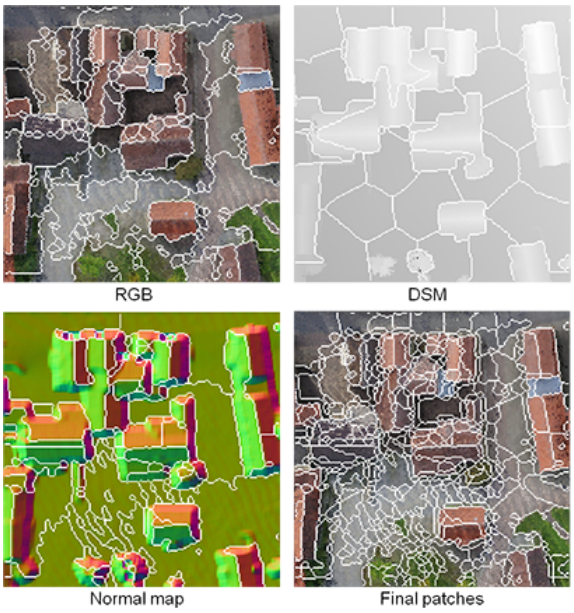

Figure 7. Patches generated by oversegmentation employing color, elevation and normal vector information

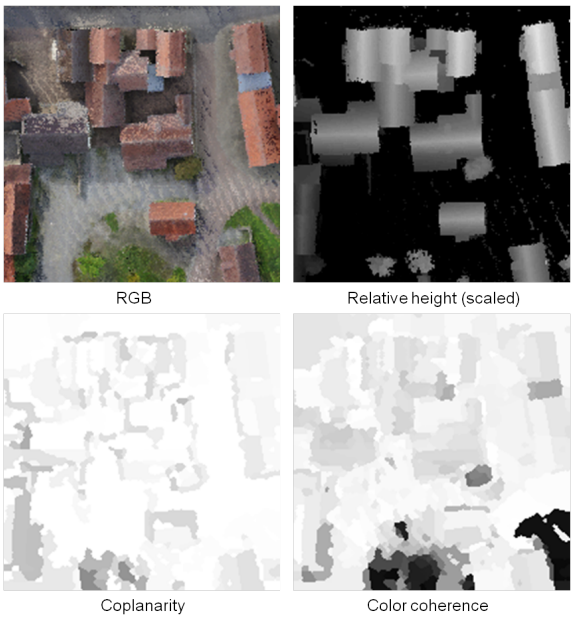

Figure 8. Relative feature of both color and geometry

e.g., diverse sizes and types of buildings and materials / colors of streets and (2) feature similarity of the objects from different classes, e.g., heights, even relative heights of buildings and trees. The challenge is to extract more inter-class stable and intra-class discriminative features from both color and geometrical information. As demonstrated in Figure 8, we employ the following synthetic features following (Huang and Mayer, 2015): (1) Relative height derived from the locally estimation ground level, (2) coplanarity used to measure how well the current point and its neighbors form a plane, which is calculated as the percentage of inliers for the common plane estimated using RANSAC (Fischler and Bolles, 1981), and (3) color coherence indicating the color difference to a reference class (vegetation), which is quantified by their distance in the $L^*a^*b^*$ space.

The features are furthermore extended by integrating the probabilities derived in Section 3 as the belief in the individual data points. In the proposed scheme, a patch of 3D points is the unit of processing. Since a patch can only be represented by a single color, the accuracy of this color is important. Instead of using an averaged value, we only keep the color information above an empirically determined threshold of 0.8 for the beliefs and calculate the representative color with the beliefs as weights. The same idea is used for the calculation of relative height. For coplanarity, the only difference is that all the data points are kept, because in this case lower probability does not mean incorrect data and all points are needed for the consensus.

4.3 Classification

A standard random forest classifier (Breiman, 2001) is employed. The calculation of features and the classification with the trained classifier are implemented aiming at parallel processing. Please note that the superpixel-based segmentation is in principle not suitable for parallelization and requires computational effort related exponentially to the image size. With the assumption that the proposed "relative" features are robust in various scenarios, which implies a generally trained classifier can be directly applied on all data partitions without additional local or incremental training, the whole scene is divided into smaller partitions which are processed separately. As long as the partitions have a reasonable size, i.e., are large enough to contain main objects like buildings and roads, the division will only marginally disturb the results. Due to the oversegmentation, the additional boundaries have only a minimum effect on the results.

A post-processing is conducted to correct trivial errors caused by data artifacts and improves the plausibility of the results. It works as a blob filter based on consistency constraints for roads and buildings.

5. EXPERIMENTS

For the village dataset (see Figure 1), we acquired a set of 296 images from the village Bonnland, Germany, flying a UAV 150 m above ground. The UAV carries a Sony ILCE-7R camera with a fixed-focus lens having a focal length of 35 mm. Hence, the images with a size of 7360×4912 pixels have a ground resolution of approximately 2 cm. Each image overlaps with ten other images on average. The scene shows buildings with paved ground between them in its center and forest and grass around them.

The experiments are performed on a stand-alone standard PC (dual socket) with two Intel Xeon processors (135 W, 3.5 GHz) with 16 cores in total. Furthermore, the computer is equipped with a NVidia GeForce GTX 970 graphics card and a Virtex-6 Board for the Field Programmable Gate Array (FPGA). The graphics card is used in the registration step for SIFT feature extraction employing the approach of (Wu, 2007). We perform the semi-global matching (SGM) on the FPGA.

First, we derived a 3D-reconstructed scene based on images downscaled to half size. Then, we repeated the test on the images with quarter resolution. This allows a faster processing with SGM and one additionally gets rid of high frequency noise. SGM on full resolution images leads to less dense disparity maps. Furthermore, the full resolution is not necessary for our application. Both tests were performed using SGM in a multi-view configuration. The images were registered by employing (Mayer et al., 2011) in 31 minutes (half size) or 36 minutes (full size). The

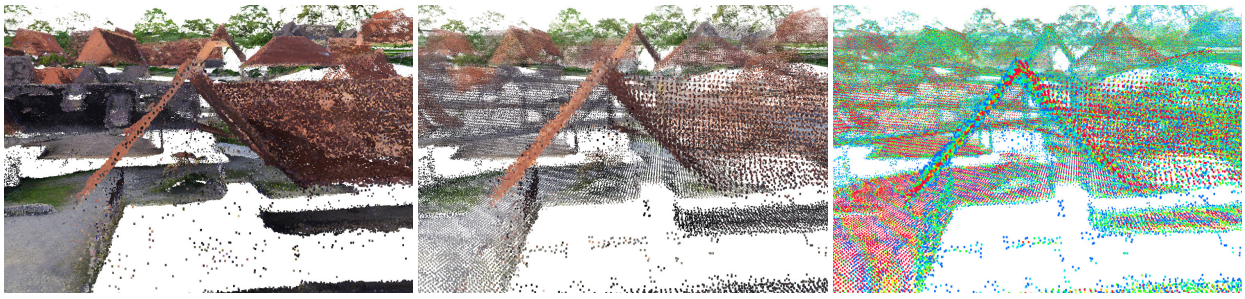

Figure 9. 3D point clouds from the village dataset (Figure 1). Left: Dense and accurate 3D point cloud derived by volumetric fusion of truncated signed distance functions (Kuhn et al., 2014). Center:Resulting point cloud from the novel fast fusion method. Right: Coded probabilities from the occupancy grid from our fusion. It is obvious, that our fusion leads to higher noise and less dense point clouds than (Kuhn et al., 2014). Nonetheless, e.g., the border of the roof is clearly visible when considering the high probabilities (red points).

multi-view SGM was performed in 148 minutes (quarter size) and in 851 minutes (half size).

The disparity maps can be transformed to 3D point clouds considering the registration of the image set. Overall, the SGM disparity maps in half resolution of our village dataset lead to 1.082.124.721 (over a billion) 3D points. The quarter resolution maps result in 406.819.206 (nearly half a billion) 3D points. After decomposition of the individual disparity maps, the point cloud is reduced to 45.319.881 (quarter resolution) and 32.136.740 (half resolution) 3D points. The decomposition size was set to an accuracy of $20\,cm$. Hence, the decomposed point clouds from half and quarter resolution disparity maps are quite similar and in the following we only consider the quarter resolution pipeline as it is much faster and generates point clouds with less holes. For the occupancy grid fusion we use a logprob $l = 1.0$ for individual points corresponding to an inlier probability of $p = 0.73$.

Because of the strong parallelization possible with 16 Central Processing Unit (CPU) cores and fast reading and writing capabilities on a RAID 0 SSD hard disk, the point cloud decomposition of the entire set of disparity maps can be processed in only 20 seconds. The fusion of the set of point clouds into an occupancy grid takes 30 seconds on a single CPU. This part can be further parallelized. Yet, because of the high implementation effort necessary, this is beyond the work of this paper. Overall, the entire fusion process was conducted on a single PC in below one minute for the village dataset.

For the evaluation of the quality of the scene classification it is important to compare the results to state-of-the-art methods. Unfortunately, for image-based reconstruction of point clouds there is no ground truth data available. For a comparison, we, therefore, compare our results to results of high-quality fusion methods. Especially, 3D reconstruction methods based on the fusion of Truncated Signed Distance Functions (TSDF) (Curless and Levoy, 1996) have been shown to produce highly accurate point clouds also for large datasets (Fuhrmann and Goesele, 2011, Fuhrmann and Goesele, 2014, Kuhn et al., 2013, Kuhn et al., 2014). In the field of large scale urban environments, particularly the work of (Kuhn et al., 2014) has been used to produce results also for urban scene classification (Huang and Mayer, 2015). Hence, we use this method to produce dense and accurate point clouds, which can be employed for a comparison to our fusion method.

The fusion of TSDF requires much memory. Hence, in (Kuhn et al., 2014) the entire space is split into subspaces which are merged subsequently (Kuhn and Mayer, 2015). To generate an accurate 3D point cloud, the fusion is done in probabilistic space where single measurements are represented by a couple of voxels. The complex fusion and high memory requirements mean that the runtime is the highest of all components when integrating it in our pipeline. E.g., for the large dataset with quarter resolution, the fusion needs more than four hours. In (Kuhn et al., 2014) an additional meshing of the resulting point cloud is proposed, which was not used in our experiments. Compared to this our novel fusion method is about 250 times faster and even though the result is much noisier and less dense (see Figure 9), it produces a similar scene classification (cf. Figure 10).

The Bonnland data cover about 0.12 square kilometer of undulating terrain. The classification is performed on the rasterized point cloud with a resolution of 0.2 meter. Figure 10 shows selected results of the classification. Performance on the datasets with conventional (non-probabilistic) and the proposed probabilistic fusion method are presented for comparison. We define the four classes of object: Ground (gray), building (red), high vegetation (trees, green), and low vegetation (grass, blue). The runtime for the whole area is about 17 minutes. 51.2% of the time is applied for the oversegmentation and the rest for the feature calculation (4.6%) as well as the classification (44.2%), which are processed in parallel with the above mentioned hardware setup. The data have been divided into 48 tiles of 200×200 pixels/data points (examples are shown in Figure 10). The partitioning significantly reduces the runtime of the oversegmentation, which uses global optimization and is thus the most time-consuming part. The precondition is, as mentioned in Section 4.3, that the features are robust against variations of both ground height and color appearance so that a generic classifier can be employed for all tiles. The latter also means the tiles could be calculated in parallel on a computer cluster, although in this paper they are processed sequentially on the stand-alone PC, where the given runtime has been measured.

Examining Figure 10, it is obvious that scene classification of the noisy (fast fused) point cloud is only accurate when considering appropriately determined probabilities. Without these probabilities the classification produces significant errors (e.g., middle right image). By means of probabilities the quality of the results is similar to classification considering complex fusion. In a small region in the top part of the images in the right column the courtyard was classified correctly only based on complex fusion results. In the middle column it is obvious that vegetation was best classified considering fast probabilistic fusion. Furthermore, the buildings are better separated from the surrounding in all examples for fast probability fusion. This can especially be seen at the standalone building in the middle column.

Figure 11 presents a difficult case with results of limited quality. The proposed patch-wise scheme provides a time-efficient classification, but the performance might be affected by an incorrect oversegmentation (cf. Section 4.1). The latter is mostly

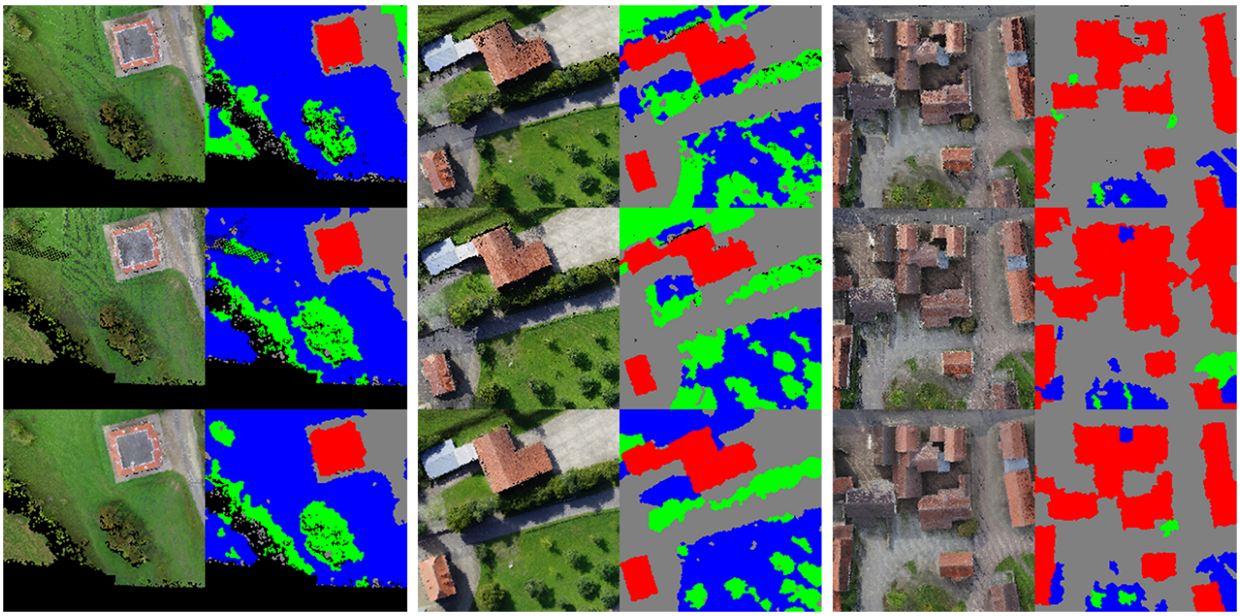

Figure 10. The top row shows classification results on the point cloud from complex fusion (Kuhn et al., 2014) (Figure 9 [left]). Results for the fast proposed probabilistic fusion but without considering probabilities (Figure 9 [centre]) are shown in the middle row and for probabilities derived by Equations 1 to 3 (Figure 9 [right]) results are given in the bottom row (red – buildings, gray – ground, blue – grass, and green – trees).

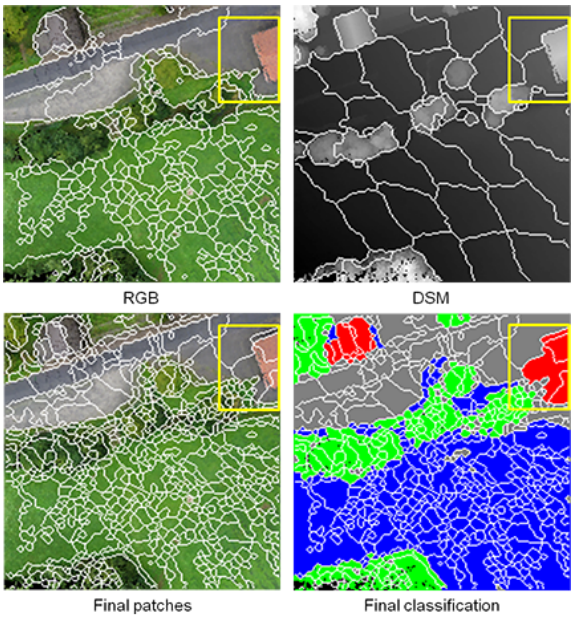

Figure 11. Classification results affected by artifacts in the data (red – buildings, gray – ground, blue – grass, and green – trees)

caused by artifacts of the data points concerning color or geometry. As shown in Figure 11 (top), the partial roof at the top-right corner contains false points on the edges, which result in a false segmentation concerning both color and DSM. Once a patch is falsely segmented crossing two classes, this error directly affects the final result.

In summary, we have shown that an accurate scene classification from unordered images is possible with our method for a dataset within three and a half hours on a single PC. Table 1 gives an overview of the runtime for the components of our processing pipeline for quarter and half resolution images.

	IR	SGM	PCF	SC	Overall
half res	36	851	1	17	912
quarter res	31	148	1	17	204

Table 1. Runtime in minutes of Image Registration (IR), Semi-Global Matching (SGM), Point Cloud Fusion (PCF) and Scene Classification (SC). The input in the first row are the images in half resolution (res), while in the second row the images in quarter resolution are used. The overall runtime for the presented results (Figures 1 and 10) is 204 minutes.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a method for fast fusion of disparity maps in 3D space. It complements state-of-the-art disparity map fusion as it is much faster and yields a better quality for scene classification. The method can cope with outliers as they are probabilistically considered. The fast fusion is very useful for applications that have no need for dense point clouds derived from high resolution images. We have shown that occupancy grids based on octrees are suitable for this task.

We have also proposed to employ a supervised classification based on color and elevation data with (1) robustness against heterogeneous appearance of objects and variable topography and (2) time-efficient patch-wise feature extraction and classification.

We are aware that an incorrect oversegmentation caused by data artifacts is one of the main sources of error. Besides improved data quality, a post-processing is considered to context-sensitively fill unavoidable gaps in the data. Furthermore, we consider to extend the class definition with additional and/or refined classes such as cars, water bodies and different roof types.

An important benefit of our method is that already in disparity map generation the number of points can be increased by limiting outlier filtering. The above results show that outliers can be classified based on the probabilities estimated in the occupancy grid fusion. In MVS estimation by SGM, multiple dis-

parity maps from image pairs are fused. Disparities with insufficient correspondences in n views are filtered. On one hand, this leads to stable 3D points, but on the other hand, important points may be filtered leading to gaps. Keeping instable points leads to more complete point clouds and, hence, could further improve the scene classification.

REFERENCES

Breiman, L., 2001. Random forests. Machine Learning 45(1), pp. 5–32.

Curless, B. and Levoy, M., 1996. A volumetric method for building complex models from range images. In: 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 303–312.

Ernst, I. and Hirschmüller, H., 2008. Mutual information based semi-global stereo matching on the GPU. In: 4th International Symposium on Advances in Visual Computing, pp. 228–239.

Felzenszwalb, P. and Huttenlocher, D., 2004. Efficient graph-based image segmentation. International Journal of Computer Vision 59(2), pp. 167–181.

Fischler, M. and Bolles, R., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM 24(6), pp. 381–395.

Fuhrmann, S. and Goesele, M., 2011. Fusion of depth maps with multiple scales. In: Proceedings of the 2011 SIGGRAPH Asia Conference, pp. 148:1–148:8.

Fuhrmann, S. and Goesele, M., 2014. Floating scale surface reconstruction. In: SIGGRAPH Conference, pp. 46:1–46:11.

Guo, L., Chehata, N., Mallet, C. and Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. ISPRS Journal of Photogrammetry and Remote Sensing 66(1), pp. 56 – 66.

Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2), pp. 328–341.

Hirschmüller, H., 2011. Semi-global matching - motivation, developments and applications. In: Photogrammetric Week '11, pp. 173–184.

Hirschmüller, H. and Scharstein, D., 2009. Evaluation of stereo matching costs on images with radiometric differences. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(9), pp. 1582–1599.

Hoberg, T., Rottensteiner, F., Feitosa, R. Q. and Heipke, C., 2015. Conditional random fields for multitemporal and multiscale classification of optical satellite imagery. IEEE T. Geoscience and Remote Sensing 53(2), pp. 659–673.

Huang, H. and Mayer, H., 2015. Robust and efficient urban scene classification using relative features. In: 23th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, Seattle, WA, USA, pp. 81:1–81:4.

Huang, H., Jiang, H., Brenner, C. and Mayer, H., 2014. Object-level segmentation of rgbd data. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences II-3, pp. 73–78.

Kuhn, A. and Mayer, H., 2015. Incremental division of very large point clouds for scalable 3D surface reconstruction. In: ICCV Workshop on Multi-Sensor Fusion for Dynamic Scene Understanding.

Kuhn, A., Hirschmüller, H. and Mayer, H., 2013. Multi-resolution range data fusion for multi-view stereo reconstruction. In: 35th German Conference on Pattern Recognition, pp. 41–50.

Kuhn, A., Mayer, H. and Hirschmüller, H., 2014. A TV prior for high-quality local multi-view stereo reconstruction. In: 2nd International Conference on 3D Vision, pp. 65–72.

Mayer, H., 2015. From orientation to functional modeling for terrestrial and UAV images. In: Photogrammetric Week '15, pp. 165–174.

Mayer, H., Bartelsen, J., Hirschmüller, H. and Kuhn, A., 2011. Dense 3D reconstruction from wide baseline image sets. In: 15th International Workshop on Theoretical Foundations of Computer Vision, pp. 285–304.

Moore, A., Prince, S., Warrell, J., Mohammed, U. and Jones, G., 2008. Superpixel lattices. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1–8.

Moravec, H. and Elfes, A., 1985. High resolution maps from wide angle sonar. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 105–112.

Niemeyer, J., Rottensteiner, F. and Soergel, U., 2013. Classification of urban lidar data using conditional random field and random forests. In: Urban Remote Sensing Event (JURSE), 2013 Joint, pp. 139–142.

Ren, X. and Malik, J., 2003. Learning a classification model for segmentation. In: IEEE International Conference on Computer Vision (ICCV), pp. 10–17 vol.1.

Rothermel, M., Wenzel, K., Fritsch, D. and Haala, N., 2012. SURE: Photogrammetric surface reconstruction from imagery. In: LC3D Workshop.

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. IEEE T. Geoscience and Remote Sensing 50(11), pp. 4534–4545.

Schmidt, A., Niemeyer, J., Rottensteiner, F. and Soergel, U., 2014. Contextual classification of full waveform lidar data in the wadden sea. Geoscience and Remote Sensing Letters, IEEE 11(9), pp. 1614–1618.

Thrun, S., Burgard, W. and Fox, D., 2005. Probabilistic Robotics (Intelligent Robotics and Autonomous Agents). The MIT Press.

Ummenhofer, B. and Brox, T., 2015. Global, dense multiscale reconstruction for a billion points. In: IEEE International Conference on Computer Vision (ICCV), pp. 1341–1349.

Vosselman, G., 2013. Point cloud segmentation for urban scene classification. ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-7/W2, pp. 257–262.

Wu, C., 2007. SiftGPU: A GPU implementation of scale invariant feature transform (sift). In: Online.

Wu, C., 2013. Towards linear-time incremental structure from motion. In: International Conference on 3D Vision, pp. 127–134.

Wu, C., Agarwal, S., Curless, B., and Seitz, S. M., 2011. Multicore bundle adjustment. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3057–3064.

Zhang, X., Zang, A., Agam, G. and Chen, X., 2014. Learning from synthetic models for roof style classification in point clouds. In: 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, New York, NY, USA, pp. 263–270.