FUSION OF FEATURE BASED AND DEEP LEARNING METHODS FOR CLASSIFICATION OF MMS POINT CLOUDS

D. Tosic[1,2], S. Tuttas[2], L. Hoegner[1], U. Stilla[1]

[1] Photogrammetry and Remote Sensing, Technical University of Munich (TUM), Germany
- (dragana.tosic, ludwig.hoegner, stilla)@tum.de
[2] 3D Mapping Solutions GmbH, 83607 Holzkirchen, Germany
- (dragana.tosic, sebastian.tuttas)@3d-mapping.de

ICWG II/III: Pattern Analysis in Remote Sensing

KEY WORDS: Deep learning, Segmentation, Urban scene, Classification, Point clouds, Mobile mapping, HD Maps

ABSTRACT:

This work proposes an approach for semantic classification of an outdoor-scene point cloud acquired with a high precision Mobile Mapping System (MMS), with major goal to contribute to the automatic creation of High Definition (HD) Maps. The automatic point labeling is achieved by utilizing the combination of a feature-based approach for semantic classification of point clouds and a deep learning approach for semantic segmentation of images. Both, point cloud data, as well as the data from a multi-camera system are used for gaining spatial information in an urban scene. Two types of classification applied for this task are: 1) Feature-based approach, in which the point cloud is organized into a supervoxel structure for capturing geometric characteristics of points. Several geometric features are then extracted for appropriate representation of the local geometry, followed by removing the effect of local tendency for each supervoxel to enhance the distinction between similar structures. And lastly, the Random Forests (RF) algorithm is applied in the classification phase, for assigning labels to supervoxels and therefore to points within them. 2) The deep learning approach is employed for semantic segmentation of MMS images of the same scene. To achieve this, an implementation of Pyramid Scene Parsing Network is used. Resulting segmented images with each pixel containing a class label are then projected onto the point cloud, enabling label assignment for each point. At the end, experiment results are presented from a complex urban scene and the performance of this method is evaluated on a manually labeled dataset, for the deep learning and feature-based classification individually, as well as for the result of the labels fusion. The achieved overall accuracy with fusioned output is 0.87 on the final test set, which significantly outperforms the results of individual methods on the same point cloud. The labeled data is published on the TUM-PF Semantic-Labeling-Benchmark.

1. INTRODUCTION

1.1 Motivation

Increasing need for fast and accurate 3D spatial data (e.g. for designing HD maps for autonomous driving) has led to rapid development of Mobile Mapping Systems (MMS) in terms of accuracy and scanning density, which further enabled extensive research in the topic of 3D scene semantic classification. The course of development of MMS is thoroughly described regarding different aspects of the technology in several recent reviews (Tao , Li, 2007, Puente et al., 2013).

Our work offers a solution for semantic classification of outdoor-scene point clouds by utilizing combination of feature-based approach for semantic segmentation of point clouds with deep learning approach for semantic segmentation of images. The major goal is an output with enhanced classification accuracy, compared to the outputs of individual methods applied for the same task.

For these purposes, two types of classification are performed upon data collected with an MMS (Figure 1): 1) Feature-based approach is applied as in (Sun et al., 2018, Xu et al., 2018): firstly, point cloud is organized into a supervoxel structure for capturing geometric characteristics of points, followed by defining local context for each supervoxel for gaining contextual information. Secondly, several geometric

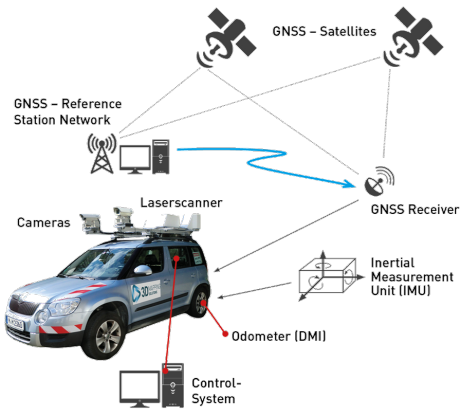


Figure 1. High precission mobile mapping system and sensors used (3D Mapping Solutions GmbH)

features are extracted for appropriate representation of the local geometry, followed by removal of the effect of local tendency for each supervoxel in order to enhance distinction between similar structures. And lastly, Random Forests (RF) algorithm is applied for assigning labels to supervoxels and points within them. 2) Deep learning approach is employed through semantic segmentation of MMS images of the same scene. To achieve this, Pyramid Scene Parsing Network (Zhao et al., 2017) is used

and resulting segmented images with each pixel containing a class label are then projected into the point cloud, enabling classification for each point. At the end, fusion of the point clouds from the same urban scene, classified with these two methods is presented as experiment result and the performance of our method evaluated on a manually labeled dataset.

1.2 State of the art in classification of laser scanning point clouds

To provide semantic infomation from data acquired by MMS, different methods for data classification, segmentation and object recognition are developed over the time and described in the literature (Guan et al., 2016, Ma et al., 2018). According to (Mei et al., 2018), these methods can rougly be devided into three groups: feature-based methods, deep learning methods and semi-supervised learning methods.

Achievements of deep learning methods without hand-crafted features, especially Convolutional Neural Networks (CNN) in image segmentation (Garcia-Garcia et al., 2017) inspired similar techniques for 3D point clouds classification. Several suggested methods perform segmentation by feeding a designed CNN with 3D tensors (Lai et al., 2014). The most significant challenges of 3D deep learning, such as large computation time and increased chance of overfitting due to the costly and therefore limited training samples have led to an alternative strategy of applying neural networks to 2D tensors, gained by projecting the point clouds onto a 2D image plane. Segmentation is then performed on such images and predicted labels are assigned to points in a point cloud via back-projection (Lawin et al., 2017, Boulch et al., 2018).

2. METHODOLOGY

The applied methodology consists of five major parts. 1) In the first part the feature-based method (Method 1) is applied and supervoxel-based classification is gained. In order to obtain point-wise classification, the label of the nearest supervoxel is assigned to each point. 2) Deep learning method (Method 2) is employed through semantic classification of images from three MMS cameras via neural network PSPNet. 3) Evaluation of both methods individually is done against the first part of the test data (Test 1) and performance of each method regarding each class are considered for the next step. 4) The second part of the test data (Test 2) was also individually classified by each of the methods. However, the decision about the class labels is made upon the performance of each method in the evaluation against data in Test 1. In this way, a certain amount of unbiasedness is achieved. 5) At the end, an evaluation of the finally classified Test 2 data is done and the results are compared to the results of the individual methods for the same point set. An overview of the method is given in the Figure 2 and elaborated regarding individual methods in the following subsections.

2.1 Feature-based point cloud classification

Firstly, a method for classification of the laser scanning point clouds with previous over-segmentation via supervoxel-structures is used as proposed by (Sun et al., 2018). To achieve 3D partitioning in the form of supervoxels, voxel seeding is performed within a regular grid and these are considered as centers of the supervoxels. Then, the connections of voxels are estimated and used as condition for grouping

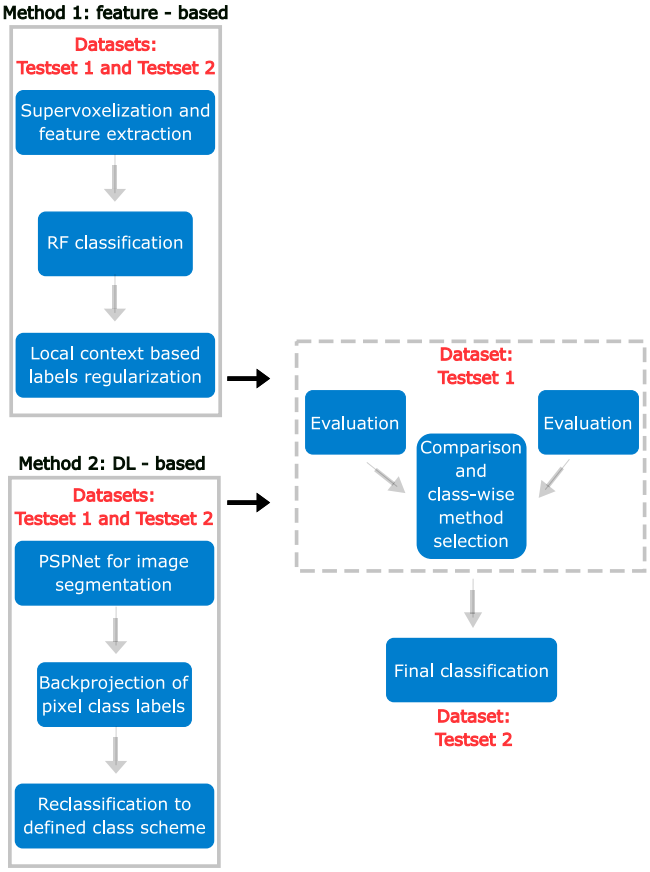


Figure 2. Workflow of the proposed method

them together within certain neighborhoods. The connectivity is estimated by calculating the distance D in feature space as (Sun et al., 2018):

$$D = \sqrt{w_c D_c^2 + w_s \frac{D_s^2}{R_{seed}^2} + w_n D_n^2}\,, \tag{1}$$

considering D_c, D_s and D_n - the distances in Euclidean, color and normal spaces, respectively, whereas w_c, w_s and w_n are the weighting factors. However, in this work only spatial distance and normal vectors are considered as a criteria for creating supervoxel structures. Since this method utilizes geometric features, an appropriate representation of local geometry is necessary. For this purpose, 3D shape features are introduced, first by deriving respective eigenvalues λ_i, $i \in \{1, 2, 3\}$, and after that by calculating linearity, planarity, scattering, omnivariance, anisotropy, eigenotropy and local curvature, as proposed in (Weinmann et al., 2015). Additional features are considered as follows: height-features, orientation features, radiometric features, as well as features gained by subtraction the local context of each supervoxel to enhance differentiation between similar structures. The latter is referred to as the detrending process, during which the local tendency of each supervoxel is calculated in feature space by considering the neighboring supervoxels and expressed through feature histogram of the local tendency V_{LT}. It is then subtracted from the feature histogram of a considered supervoxel V_S. The detrended geometric feature histogram is then obtained with:

$$V = V_S - V_{LT} \tag{2}$$

The final feature histogram V_F is obtained by weighted combination of V and V_S:

$$V_F = \{V_S, k \cdot V\} \tag{3}$$

where weight k is a weight for each local tendency, estimated by the number of supervoxels in the local context. After geometric features are calculated for all supervoxels, supervised classification with RF algorithm is performed (Breiman, 2001) to assign semantic labels to supervoxels and points within them, based on the extracted features.

2.2 Image segmentation

In our work, the approach of gaining point-cloud images via backprojection and feeding these into a neural network, encountered in related work, is omitted. Instead, image data from three MMS cameras are segmented with Pyramid Scene Parsing Network (PSPNet) (Zhao et al., 2017), which successfully copes with one of the greatest challenges of FCNs - capturing of the global scene context and including that information in class prediction. For this purpose, PSPNet incorporates a pyramid pooling module, as one of the main contributions of that proposed architecture. In order to extract the feature map of the input image from the last CONV layer of a CNN, PSPNet uses pretrained ResNet (He et al., 2016) with dilated strategy (Yu , Koltun, 2015) for the receptive field expansion. The feature map is passed further as the input to the pyramid pooling module, in which scene context information is collected via four levels of a pyramid. Each of the pyramid levels are pooling kernels of different sizes: 1×1, 2×2, 3×3 and 6×6, respectively. In that way, features from different sub-regions are gathered on each pyramid level. Global prior information, captured from pyramid levels, is fused with the original feature map, yielding an improvement compared to global pooling (Liu et al., 2015) in capturing the scene context and incorporating that information in image segmentation process.

Class-assignment from labeled pixels to each point in the point cloud is enabled through the precise calibration of MMS sensors. Therefore, relative position of all cameras and laser scanners on the used platform (Figure 1) are known and they operate synchronized in time. This enables precise matching of images and scans of the same scenery. Following time synchronization, a projection

$$f : R^3 \rightarrow R^2 \tag{4}$$

is performed, by calculating homogeneous coordinates for each 3D point of the current scene for the purpose of right pixel-label assignment.

$$X_c = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} X_w \tag{5}$$

$$x = K[I|0]X_c \tag{6}$$

where K is the camera calibration matrix, X_c, X_w are point coordinates in camera and world coordination system, respectively and x are pixel coordinates of each projected 3D point. For determination of laser scanning points visible from each camera position, Hidden Points Removal is applied (Katz et al., 2007).

2.3 Fusion of classification outputs

In this section a further elaboration of the fusion step is provided and illustrated in Figure 3. Succeeding point-wise classification of the point cloud with both mentioned methods (Figure 2), our final result is a fusion of two classified point clouds, which yields a possibility for a wide field of analyses.

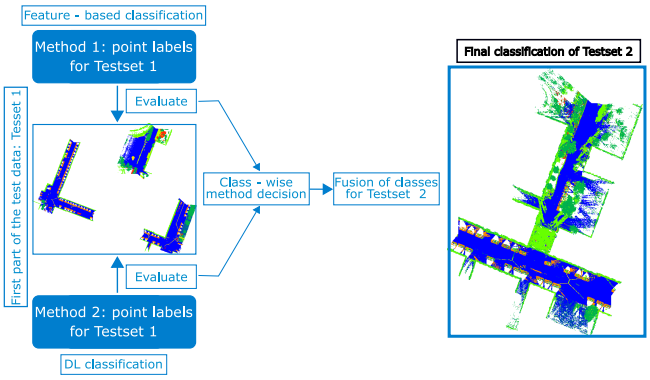


Figure 3. Overview of the method for fusion of feature-based and DL classification

Obtaining the final, fusioned output involves four major parts. 1) Firstly, classification outputs from a) the feature-based method (Method 1) and b) the deep learning method (Method 2) are obtained individually, by applying the methodology described in previous sections. 2) Following that, an evaluation of the two methods is done against the Testset 1 and based on their performance, one of the two methods is chosen to deliver labels for each of the classes in the final step. 3) Afterwards, the classification of the Testset 2 by each method individually is done and finally 4) the class labels are chosen according to performance of each method in step 2. In case of contradiction, e.g. Method 1 voted for the class "vegetation", while for the same point a class "building" is assigned by the Method 2, the preferred is the label of a class which has had higher F_1 score in the evaluation against the data in Testset 1. The final output is the result of the labels fusion on the Testset 2. This output is then evaluated against the ground truth data and the result of evaluation is compared to the individual results of Method 1 and Method 2 on the same dataset. The reliability of the predicted class labels is assessed through redundancy of resulting labels considered for each point. Segmentation results from images of different cameras are compared and weighted against each other and against the results of the feature-based method. It is important to note that the comparison of the labels achieved with Method 1 and Method 2 is only done for the points considered by both methods, which are the points visible from MMS cameras. This is an example of how feature-based method compensates for the limitations of DL-based one.

3. EXPERIMENTS

3.1 Datasets

Data acquisition for the practical part of this thesis was done with the high-precision MMS from 3D Mapping Solutions GmbH (Gräfe, 2007, Gräfe, 2009) (Figure 1) in the area of about 50 000 m^2 around the Technical University of Munich - in Gabelsbergerstrasse, Arcistrasse, Theresienstrasse and Luisenstrasse. The two laser scanners operate with a frequency of 200 profiles in a second and with the repetition rate accuracy for each point of approximately 0.5 mm.

Figure 4. Examples of images recorded with MMS front cameras

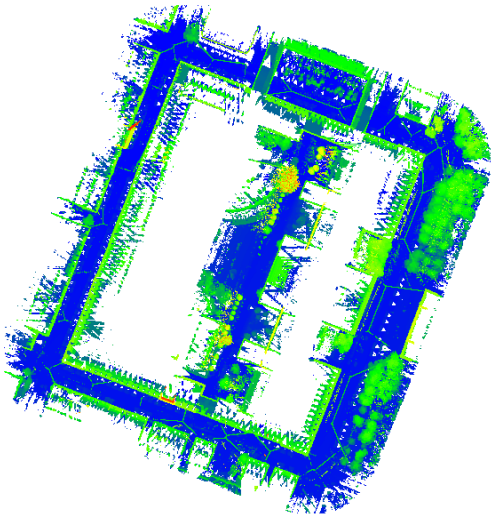


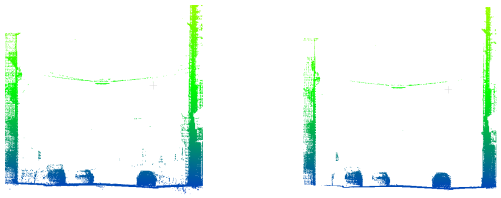
Figure 5. Nadir view over the complete point cloud dataset: raw data colored by height

Such performances enable very high resolution of the point clouds, with an approximate resolution of 2500 points per m^2 within a distance of $3\,m$, assuming the driving speed of $72\,km/h$ (Gräfe, 2018). As a result of lower driving speed, the point density of the dataset used for experiments in this thesis was around 14.500 points per m^2 within a distance of $3\,m$. Due to the rotated position of the scanners relative to each other, a certain time difference in recording the same point is present. All sensors are calibrated and co-registered into a common coordinate system with the center in the IMU.

The final dataset consisted of 1299 images from three industrial RGB cameras, two of which were mounted on the front of the vehicle (resolution 2336×1776 px) and one on the back (resolution 2432×2058 px). In total, around 320 million laser scanning points were acquired in the area by the two line-scanners. Figure 5 provides an outlook over the collected point cloud of the entire area and examples of images from MMS front cameras are provided in Figure 4.

Before proceeding with data processing for the purposes of experimental work, data was cleaned from outliers by applying noise suppressing and manual cleaning. A comparison of data before and after outliers removal is shown in Figure 6. Significant amount of data is still kept, as well as original density, with major intention to provide convenient base for HD-Maps generation.

Further steps were generation of ground truth and data partitioning into train and test partition as described in following sections.



(a) (b)

Figure 6. Rendered by height: (a) Raw data with noise (b) Cleaned data

3.2 Generation of ground truth

In order to achieve precise evaluation of the classification and for the purposes of training the classifier, an accurate manually labeled point cloud for the whole dataset was created as ground truth. The points are assigned with unique labels from the selected ten semantic classes to describe different objects in urban area: *man-made terrain, natural terrain, vegetation, building, hard scape, pole-object (traffic signs, traffic lights included), bicycle, vehicle, man-made object and human*. The definition of classes is based on synchronization between Cityscapes (Cordts et al., 2016) and ETH Semantic3D benchmark (Hackel et al., 2017). After the careful annotation, the entire laser scanning point cloud was divided into train and test part and 50% of the acquired points were used to extract the features for the training phase of the RF-classifier (Figure 7). The second half of the dataset was further divided and the result of this partitioning are two parts of the test data: one part is used for setting the class-wise weights based on the performance of each of the two methods individually (shown in Figure 8) and the second part is used for evaluation of the final output of the fusion (in Figure 9).

3.3 Results and discussion

Method 1: Feature-bassed method In the first part of the experiment, the methodology described in Section 2.1 is put into practice. For the segmentation part, the voxel size is finally set to $0.2\,m$ and the seed resolution (distance between seeding voxels) for supervoxelization is set to $0.6\,m$. Such resolutions are chosen with two main intentions: 1) to gain neighborhoods large enough to capture the context information and 2) to obtain satisfying seed resolution for nearest-neighbor point-wise labeling which followed at the end by finding the nearest supervoxel of each point. In the classification phase, the number of trees for training the RF classifier is set to 200, by observing out-of-bag error in the training stage. Furthermore, 50% of the points in original point cloud were used for training, while only Testset 1 (around 25% of the total number of points) was used to test the classification and set the weights for the fusion.

Figure 10 provides the visual result of the feature-based classification of the points belonging to the Testset 1. As visible, there is a large area of man-made terrain, misclassified as building (enclosed with black quadrilateral). The probable reasons for such misclassification are that the class man-made terrain is trained on the area around the campus with only streets represented as this class. Therefore, the different context of the area inside of the campus - mostly concrete paver blocks as opposed to asphalt on the streets, was the probable cause of misclassification. Similarly, due to the similar geometrical characteristics and probably because of relatively low voxel

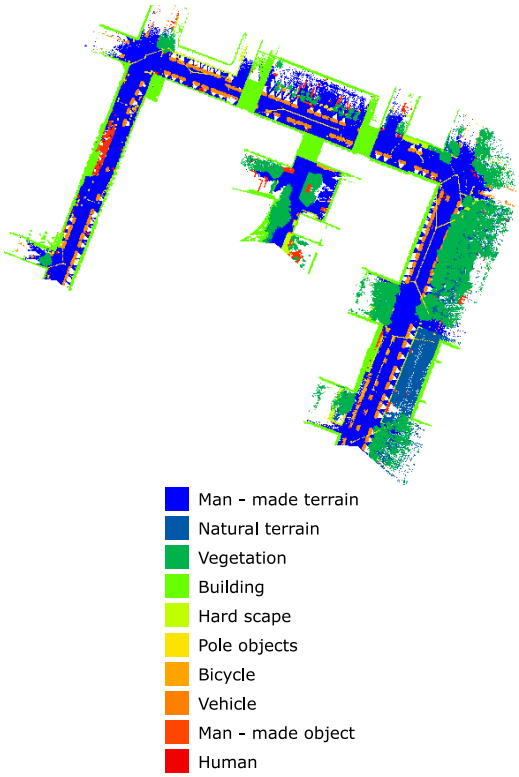


Figure 7. Training data used for training the RF classifier: part of the Luisenstrasse, Theresienstrasse and part of the Arcisstrasse

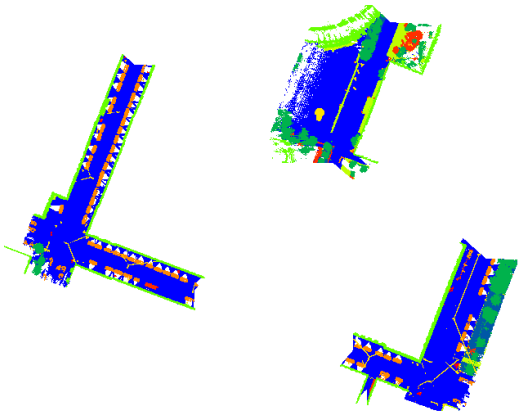


Figure 8. Testset 1: parts of Luisenstrasse and Gabelsbergerstrasse and a part of TUM inner courtyard. Used for deciding between the two methods for labels of each class. Legend as presented in Figure 7

resolution for capturing the context of smaller objects and details, parts of buildings were classified as hard scapes (Figure 11). This problem is partially solved through DL-based method, since it proves better results for most of the classes for points in camera view, as shown in Table 2.

Method 2: DL-based classification Contrary to the method in original proposal by (Zhao et al., 2017), who based their implementation on Caffe framework (Jia et al., 2014), PSPNet for image segmentation in scope of this work is based on the implementation by (Kryvoruchko et al., 2017-2019) in TensorFlow software library (Girija, 2016). The network trained on Cityscapes dataset is used and evaluation results on validation dataset are given in Table 1. Both, training imagery from Cityscapes dataset, as well as the test data provided by the

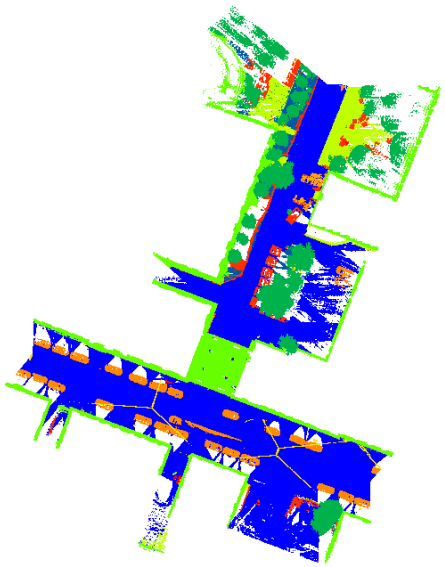


Figure 9. Testset 2: a part of the Gabelsbergerstrasse and of the TUM inner courtyard. served as the final test data for evaluating the fusion. Legend as presented in Figure 7

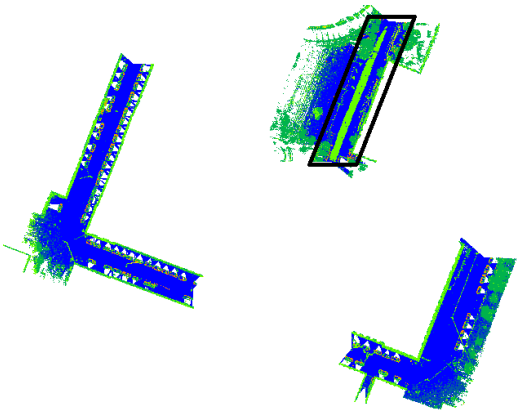


Figure 10. Result of the feature-based classification on the Testset 1

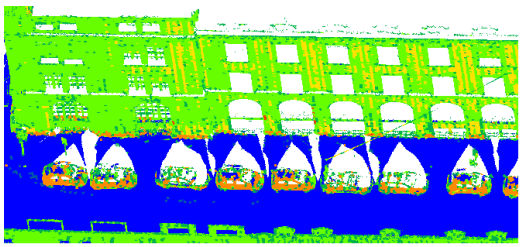


Figure 11. Misclassification of parts of building facade as hard scape (green-yellow patches)

three cameras of the used MMS system depict urban scenery and therefore, the learned parameters proved decent efficiency in the case of the testing image set.

In order to proceed with fusion of the classes and evaluation considering ground truth labels, the points are reclassified following the defined ten classes (Section 3.2) and an example for reclassified images is shown in Figure 12b.

Since a significant redundancy is achieved by providing three labels for each point visible from the cameras, the final choice

Classes	IoU class
Road	0.972
Sidewalk	0.781
Building	0.888
Wall	0.531
Fence	0.502
Pole	0.342
Traffic light	0.413
Traffic sign	0.607
Vegetation	0.880
Terrain	0.611
Sky	0.923
Person	0.611
Rider	0.302
Car	0.919
Truck	0.698
Bus	0.747
Train	0.651
Motorcycle	0.370
Bicycle	0.622
Score Average	0.651

Categories	IoU
Sky	0.917
Human	0.630
Vehicle	0.885
Flat	0.976
Object	0.432
Construction	0.882
Nature	0.885
Score Average	0.801

Table 1. PSPNet performance on Cityscapes validation set: Left: Class - wise; Right: Category - wise

of unique labels was completed by evaluating the predicted labels resulting from each of the three sets of images. This evaluation was performed against the ground truth for the Testset 1, shown in Figure 8. In this part of the test data, various objects are present, such as vegetation, classic city scapes with vehicles and buildings, as well as the part of the inner courtyard with objects particularly challenging for the classification. Based on F_1 score, which combines precision and recall, labels are combined and weighted for the second part of the evaluation. More precisely, labels for each class were chosen from the image set which proved the highest F_1 score for the corresponding class. Such labels are finally used for fusion with the feature-based method for classification. It is important to point out that these evaluation results relate only to the classified points. Since the field of view of the MMS cameras is limited, a significant amount of points is "unseen" and therefore not classified. The feature-based method, however, is able to classify all the points. Therefore DL method can only contribute to the final classification for the points visible from camera.

Fusion In this part of the experiment, final DL point labels are combined with point labels gained through feature-based classification. For this purpose, decision is made for the points considered by both methods, which are the points in the field of view of MMS cameras and classified by DL. Methods 1 and 2 are compared regarding their F_1-scores, as shown in Table 2.

Method / Class	Method 1 (Feature-based)	Method 2 (DL) (Deep Learning)
	F_1 score	
man-made terrain	0.840	**0.931**
natural terrain	**0.259**	0.080
vegetation	0.421	**0.433**
building	0.618	**0.857**
hard scape	**0.263**	0.014
pole objects	0.011	**0.321**
bicycle	0.006	**0.246**
vehicle	0.620	**0.858**
man-made object	0.031	**0.040**
human	-	**0.164**

Table 2. Comparison of F_1 scores of both methods, calculated in evaluation against the Testset 1. Bold: the highest score for each class

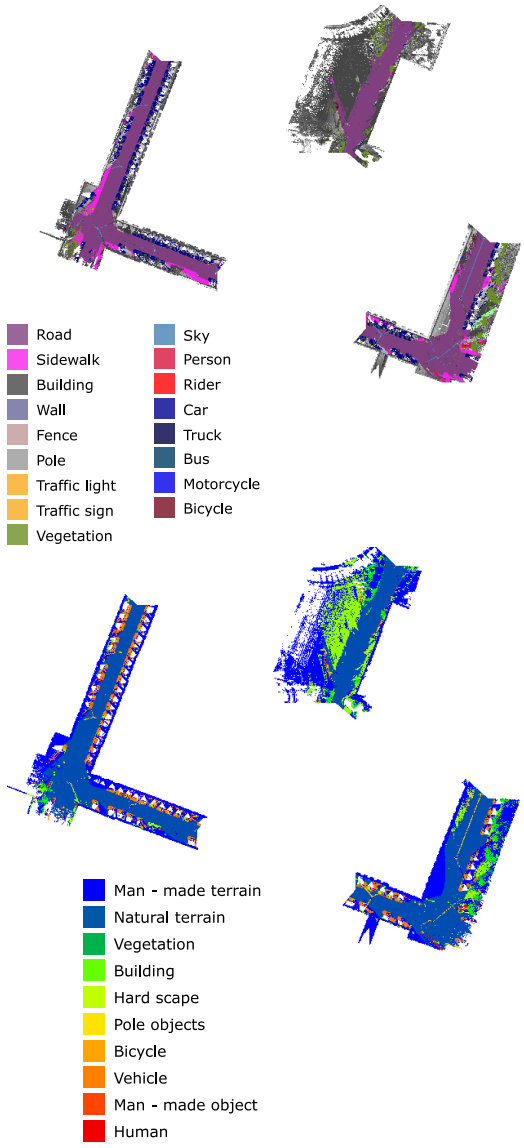


Figure 12. Classes from segmented images of the front-left MMS camera projected onto the Testset 1. Upper image shows the original classification scheme and image below point cloud reclassified for evaluation

Class	Precision	Recall	IoU	F_1-score
man-made terrain	0.958	0.762	0.738	0.849
natural terrain	0.231	0.094	0.072	0.134
vegetation	0.245	0.800	0.231	0.375
building	0.674	0.772	0.562	0.720
hard scape	0.787	0.135	0.130	0.230
pole objects	0.001	0.046	0.001	0.002
bicycle	0.078	0.002	0.002	0.005
vehicle	0.447	0.488	0.304	0.466
man-made object	0.415	0.010	0.010	0.019
human	-	-	-	-
Overall Acuracy	0.721			
Kappa	0.553			

Table 3. Evaluation of the feature-based method against the Testset 2 for all points in that dataset

Finally, based on these scores, it is decided that during the second test phase, labels predicted with Method 1 will prevail for classes natural terrain and hard scape and labels predicted with Method 2 will be assigned for classes man-made terrain, vegetation, building, pole objects, bicycle, vehicle, man-made

Class	Precision	Recall	IoU	F_1-score
man-made terrain	0.961	0.858	0.829	0.906
natural terrain	0.089	0.000	0.000	0.001
vegetation	0.376	0.096	0.083	0.153
building	0.838	0.488	0.446	0.617
hard scape	0.026	0.000	0.000	0.001
pole objects	0.336	0.160	0.122	0.217
bicycle	0.338	0.155	0.119	0.213
vehicle	0.747	0.671	0.546	0.707
man-made object	0.529	0.062	0.059	0.111
human	0.070	0.068	0.036	0.069
Overall Acuraccy	0.649			
Kappa	0.484			

Table 4. Evaluation of the DL-based method against the Testset 2 for all points in that dataset

Class	Precision	Recall	IoU	F_1-score
man-made terrain	0.955	0.957	0.916	0.956
natural terrain	0.154	0.033	0.028	0.054
vegetation	0.348	0.647	0.292	0.452
building	0.878	0.884	0.788	0.881
hard scape	0.366	0.013	0.013	0.025
pole objects	0.011	0.175	0.011	0.021
bicycle	0.329	0.156	0.119	0.212
vehicle	0.637	0.776	0.539	0.700
man-made object	0.511	0.064	0.060	0.113
human	0.070	0.068	0.036	0.069
Overall Acuraccy	0.868			
Kappa	0.776			

Table 5. Evaluation of the final output of the fusion against the ground truth of the Testset 2

object and human. The resulting classification is shown in the Figure 13 and the results after evaluating this output against the ground truth are provided in Table 5. When compared to the results of both methods individually (Tables 4 and 2), significant improvement by obtaining the fusion is evident.

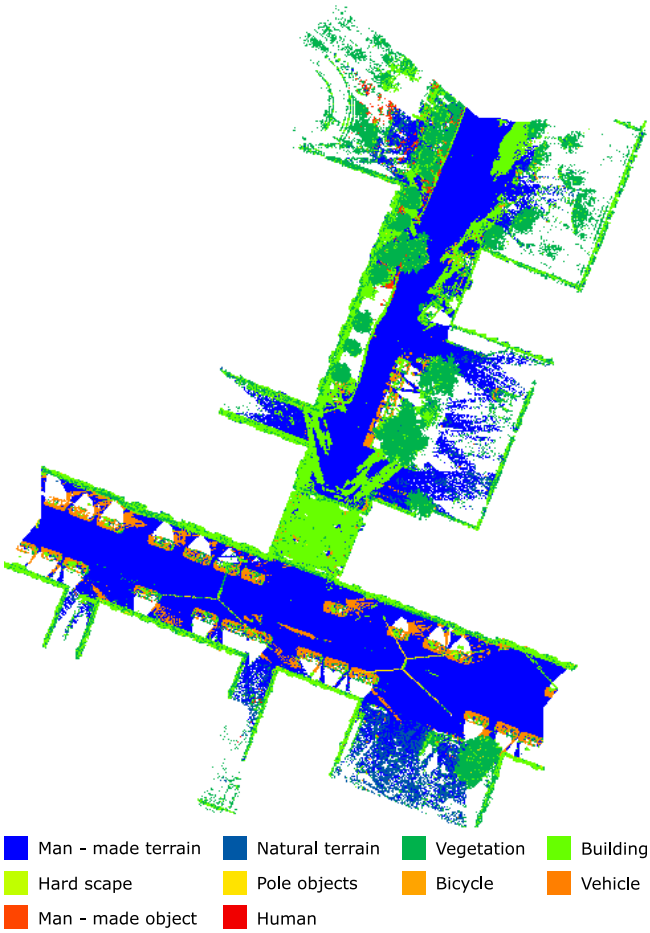


Figure 13. Final classification of Testset 2: nadir view

4. CONCLUSION

In this work, a fusion of two different methods for Mobile Mapping System (MMS) point cloud classification is proposed. A thorough class-wise comparison of the classification results between each of the methods individually and the final result of fusion is offered. Also, a significant increase in classification accuracy is demonstrated during this analysis.

Specifically in this work:

1) Classification of a highly dense point cloud is achieved, without large point reduction. Major intention behind this approach is to achieve a classification of a point cloud, which would have potential to assist the generation of HD-Maps with required spatial accuracy.

2) Experiments with supervoxel-based feature extraction and classification are utilized for point-wise labeling and influence of the local context-based labels regularization is analyzed. Such an approach reduces the computing requirements significantly, however, with this method only, the yielded classification performance is reflected through an overall accuracy of merely 0.721 on the entire Testset 2.

3) A deep learning method for semantic segmentation of images is utilized to obtain labels for points in a point cloud via backprojection of predicted pixel labels onto the point cloud. Benefiting from simpler training process and larger available training sets, 2D segmentation shows great potential also for classification in 3D. Classification of points with this method demonstrated better accuracy than the feature-based method for the points visible from camera (0.885 compared to 0.724 on the visible points in Testset 1). The limitation through the field of view is one of the major impacts in utilizing solely this method for point cloud classification, since the classification of the whole data set is impeded due to the MMS construction. Nevertheless, an overall accuracy of 0.885 on Testset 1 for the visible points, without previous fine-tuning of the neural network is quite satisfying.

4) One of the major motivations for this work was assisting the HD-Maps generation, in which extracting elements as road boundaries, buildings, vegetation and traffic signs are crucial. The experiment results on the final classification output show a possibility to assist road extraction, with the precision of 0.955 and IoU 0.916 for this class on the final test set, as well as buildings with precision of 0.878 and IoU of 0.788.

The manually labeled point cloud is published at TUM-PF Semantic-Labeling-Benchmark, under http://www.pf.bgu.tum.de/en/pub/tst.html.

REFERENCES

Boulch, A., Guerry, J., Le Saux, B., Audebert, N., 2018. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Computers & Graphics*, 71,189–198.

Breiman, L., 2001. Random forests. *Machine learning*, 45, 5–32.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena -Martinez, V., Garcia-Rodriguez, J., 2017. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*.

Girija, S.S., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *Software available from tensorflow. org*.

Gräfe, G., 2007. High precision kinematic surveying with laser scanners. *Journal of Applied Geodesy jag*, 1, 185–199.

Gräfe, G., 2009. Kinematische Anwendungen von Laserscannern im Straßenraum. PhD thesis, Univ. der Bundeswehr München, Fak. für Bauingenieur und Vermessenswesen.

Gräfe, G., 2018. Hochgenaue Qualitätssicherung für Trajektorien und deren Anwendung in Projekten der kinema-tischen Ingenieurvermessung. *DVW Bayerne.V. Gesellschaft für Geodäsiem Geoinformation and Landmanagement*, 92.

Guan, H., Li, J., Cao, S., Yu, Y., 2016. Use of mobile LiDAR in road information inventory: A review. *International Journal of Image and Data Fusion*, 7, 219–242.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M., 2017. Semantic3d.net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv: 1704.03847*.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the 22nd ACM international conference on Multimedia*, ACM, 675–678.

Katz, S., Tal, A., Basri, R., 2007. Direct visibility ofpoint sets. *ACM Transactions on Graphics (TOG)*, 26 (3), ACM, 24.

Kryvoruchko, V., Wang, C., Hu, J., Tatsch, J., 2017-2019. Pspnet-keras-tensorflow. https://github.com/Vladkryvoruchko/PSPNet-Keras-tensorflow.

Lai, K., Bo, L., Fox, D., 2014. Unsupervised feature learning for 3d scene labeling. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 3050–3057.

Lawin, F.J., Danelljan, M., Tosteberg, P., Bhat, G., Khan, F.S., Felsberg, M., 2017. Deep projective 3d semantic segmentation. *International Conference on Computer Analysis of Images and Patterns*, Springer, 9 5–107.

Liu, W., Rabinovich, A., Berg, A.C., 2015. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*

Ma, L., Li, Y., Li, J., Wang, C., Wang, R., Chapman, M., 2018. Mobile laser scanned point-clouds for road object detection and extraction: A review. *Remote Sensing*, 10, 1531.

Mei, J., Gao, B., Xu, D., Yao, W., Zhao, X., Zhao, H., 2018. Semantic Segmentation of 3D LiDAR Data in Dynamic Scene Using Semi-supervised Learning. *arXivpreprint arXiv:1809. 00426*.

Puente, I., González-Jorge, H., Martínez-Sánchez, J., Arias, P., 2013. Review of mobile mapping and surveying technologies. *Measurement*, 46, 2127–2145.

Sun, Z., Xu, Y., Hoegner, L., Stilla, U., 2018. Classification of MLS Point Clouds In Urban Scenes Using Detrended Geometric Features From Supervoxel-based Local Contexts. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4.

Tao, C.V., Li, J., 2007. *Advances in mobile mapping technology*. 4, CRC Press.

Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 286–304.

Xu, Y., Tuttas, S., Hoegner, L., Stilla, U., 2018. Voxel-based segmentation of 3D point clouds from construction sites using a probabilistic connectivity model. *Pattern Recognition Letters*, 102, 67–74.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.