STATISTICAL ANALYSIS OF AIRBORNE IMAGERY COMBINED WITH GIS INFORMATION FOR TRAINING DATA GENERATION

G. Häufel[1], D. Bulatov[1], P. Helmholz[2]

[1] Fraunhofer IOSB, Gutleuthausstr. 1, 76275 Ettlingen, Germany - (gisela.haeufel, dimitri.bulatov)@iosb.fraunhofer.de
[2] Spatial Sciences, Curtin University, GPO Box U1987, Perth, WA 6845, Australia - petra.helmholz@curtin.edu.au

ICWG II/III: Pattern Analysis in Remote Sensing

KEY WORDS: Training data generation, GIS data, Mean-Shift, K-Means, classification, land cover

ABSTRACT:

In recent years, the task of land cover classification from airborne image and elevation data advanced considerably due to enhanced applicability of CNNs (Convolutional Neural Networks). Nevertheless, CNNs require a huge amount of training data. Traditionally, few essential feature values, such as elevation or vegetation index, had been chosen to provide a coarse distinction of classes, but very often these values have to be adapted depending on the imagery. To improve this process, freely available GIS data are combined with spectral and spatial features (and their variations) following the *K-Means* and *Mean-Shift* algorithm. Based on cluster assignments to pixels, statistical analysis for extracting plausible values for distinguishing between land cover classes is applied. The resulting labeled databases are evaluated using ground truth data, and will form the basis for the training data required for CNNs.

1. INTRODUCTION

Land cover classification from actual airborne sensor data is important for many applications. One can mention analysis of accessibility after natural disasters, crops health monitoring in applications related to farming, estimation of climatically vulnerable zones due to urbanization, as well as the creation of virtual city models for city planners and civil engineers. This is why this topic has been extensively studied by many research facilities for several decades. The fast progress in sensor technology allows cost-saving data acquisition in temporarily regular intervals and in such a high resolution that smaller and narrower objects like pools, cars and footpath brings about certain challenges such as increasing intra-class variation and decreasing interclass variation (Bruzzone , Carlin, 2006). This means that objects belonging to the same class may have completely different appearances; for example, while a footpath in a forest has a completely different appearance an asphalted and gravel road, it may look more similar to an bare earth area somewhere else. The human brain is able to capture the context while taking into account the surrounding areas, and this is why in the last years, algorithms simulating the functionality of human brain have been developed and successfully applied.

In fact, with the advent of the deep learning techniques, in particular, Convolutional Neural Networks (CNN), tremendous progress has been achieved on land cover classification (Long et al., 2015, Maggiori et al., 2016, Marcos et al., 2018). As many authors have shown, even small objects of different appearances may be detected (Schilling et al., 2018, Li et al., 2016) and additional features, such as (relative) elevation or NDVI can be successfully integrated (Schilling et al., 2018), (Audebert et al., 2016). The aforementioned context is considered by the large receptive fields because convolutions evidently take place over large image regions. In order to reduce the computational burden, multiple pooling layers are introduced and, when it comes to labeling at original resolution, the images are rescaled using a encoder-decoder architectures or U-like connections. However, even with these tricks, the number of parameters to be estimated are extremely, high, for example in ImageNet (Krizhevsky et al., 2012), this number was, according to (Nayak, 2019), 62,378,344. It is therefore obvious that huge amounts of training data must be available and the claim that land cover classification is basically solved is only true given the fact that sufficient, quite huge amounts of training data are available. Thus, the challenge and the focus of the research are moving from architectures of nets to acquisition and management of training data. In this paper, we will exploit the OpenStreetMap (OSM, (Geofabrik, 2017)) data and other freely available data for creating densely labeled regions of land cover classes. The advantages of the strategy are that this data can be obtained in wider areas of the world and that it many cases it is correct, it allows staying flexible with different data acquisition conditions, like seasonal changes and sensor incidence angle. There are two disadvantages. Firstly, OSM data are not complete or not available at the desired level of details because of no entries for single trees and smaller grass areas. Furthermore, a certain risk exists they are incorrect due to changes in the scenery or due to the danger of manipulation by non-cooperative actors. In order to deal with it, this work will combine these OSM data with the actual sensor data. Features will be derived from the combined image and elevation data, after which clusters of plausible feature points will be stored using two statistical methods, namely Mean-Shift and K-Means.

2. RELATED WORK

Land cover classification is applied for distinguishing between man-made structures such as buildings or roads but also for differentiating between vegetation types or natural areas. For (Deng et al., 2015), land cover classification is a prerequisite for urban planning and environmental management. His main topic is a further development of the *normalized difference soil index* getting a better differentiation between bare soil and impervious surfaces, which are difficult to separate during land

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W16-111-2019 | © Authors 2019. CC BY 4.0 License.

cover classification. In (Helmholz et al., 2014), first, segmentation of those regions of a very high resolution image which contain agricultural areas is carried out using Markov Random Fields and then distinction into cropland and grassland is accomplished using the SVM classifier using a combined set of features (spectral, geometric, histogram-based, and textural) to distinguish between agricultural and non-agricultural objects. GIS-data are used to perform the coarse classification or filtering between these two steps.

In the last years, additional GIS information has been incrementally employed to support land cover classification. In (Grippa et al., 2018), OpenStreetMap data is used as additional landuse information, for segment-based landuse mapping, despite the fact that in some areas OSM data may be incomplete and incorrect. Furthermore, OSM data is also used in (Lopes et al., 2017), supporting special LCZ-based (Local Urban Zones) classes. These classes are divided into urban areas containing different building types and rural land cover synonymous with e.g. vegetation or soil areas. Similar to (Cheriyadat, 2014) and to remove classification inconsistencies, a buffer around OSM shapefiles such as roads or railways is attached. In (Fröhlich et al., 2013), satellite images are used for land cover classification which is carried out without human interaction based on segmentation and classification for each pixel. Additionally, information about relative elevation is utilized for classification using Iterative Context Forest. In (Kaiser et al., 2017), CNNs were employed using buildings and roads from OSM as training data whereby constant values were assigned to roads. Multi-source datasets concerning spectral reflectance and spectral indices of Sentinal-2 and Landsat-8, and OpenStreetMap data are fused together in (Qiu et al., 2018). This data stems from the So2Sat LCZ42 dataset and is distributed over 42 cities. Classification is performed using ResNet, whereby data of eight cities are selected for training. The best feature combinations for classification are determined, comparing the out comings and using majority voting. Finally, these results are used to classify the test city. Nevertheless, the authors mentioned two remaining difficulties. The height differences have to be considered and the class imbalance in the data was disregarded.

In (Bulatov et al., 2019), an approach for supervised land cover classification of aerial images is proposed. The training data generation was based on the extensive use of OSM data. To support a region-based approach for classification, a fast segmentation algorithm (Wassenberg et al., 2009) was applied resulting in the fusion of segments and rasterized OSM data. The segments were verified with respect to their relative elevation and averaged vegetation indices. After training data acquisition, more segment-based features were derived, labeled with the corresponding OSM classes, fed into a classifier, such as Random Forest, and postprocessed using smoothness priors typical for Markov Random Fields. The problem about filtering training data according to simple rules, such as relative elevation, is that the classifier will learn this simple rule and classify the data accordingly. Therefore, several suspicious segments were reclassified interactively while the authors of (Häufel et al., 2018), relied on more features and more statistical measures. Still, setting of thresholds had to be performed by the user.

Freely available GIS data offer the possibility to create large pools of training data or at least to perform a pre-selection. However, they are often given in the vector form and the rasterization without knowledge of e.g. road widths is carried out using heuristics like segmentation algorithms (Häufel et al., 2018, Bulatov et al., 2019) or constant values (Kaiser et al., 2017). Also, sometimes classes needed are not provided. For example, instead of *natural areas*, the user may be interested in trees, grass and bare soil regions: Roads may be further subdivided into parking lots, roadwork areas, junctions, and many others. This, additionally to the fact that not the whole image is labeled with OSM data, represents the main challenges. Nevertheless, our assumption is that all variation of relevant classes are present in rasterized shapefiles and consider it as our goal to identify characteristic clusters of features using the methods which are more typical for unsupervised classification. This will allow an extensive data bank of actual data exhibiting all those varieties the previous approaches had a problem with: illuminated and shaded instances, deciduous and coniferous trees and shrubs, and many others. To deal with the intraclass variations, we decided to utilize the Mean-Shift algorithm, because it only needs the pixel features itself and a bandwidth. The result is a set of cluster centers and labelings assigning some pixels to cluster. For a comparative result and to get a better impression of a second clustering algorithm, K-Means algorithm was chosen. Summarizing, our method is a semi-automatic approach relying on OSM-data, but we do not need to define interactively plausible feature values which are needed to distinguish between land cover classes. The aforementioned threshold values which had to be defined interactively in previous work (Häufel et al., 2018), shall be automised using this semi-automatic method.

3. METHODOLOGY

To explain our method in a nutshell, we will rasterize the available OSM data and combine the pixels sets resulted from the rasterization with the actual sensor data. We perform extraction of the typical spectra of these data using Mean-Shift algorithms and cluster them by K-Means. The procedure is visualized in Figure1.

This section is structured in the following way. We provide a brief description of the the basic tools for clustering, namely, K-Means and Mean-Shift, in section 3.1. Then, we refer in section 3.2 to OSM data preparation and extraction of those spectral and spatial features that are useful for distinguishing between land cover classes. Since we use K-Means and Mean-Shift for clustering, feature normalization must be performed, so that the normalized features are scale-invariant. The GIS mask preparation for selecting the training data for man-made (buildings and roads) and natural training classes (soil, trees, grass) are described more precisely in sections 3.3 and 3.4, respectively. Water bodies, mainly pools and a small lake, were not covered by OSM data. Thus, instead of clustering, quantile of a characteristic water index was used (see section 3.5) for labeling.

3.1 Clustering algorithms: Mean-Shift and K-Means

In this section, we provide a short overview about both applied clustering methods, their advantages and disadvantages. The K-Means algorithm due to (Kanungo et al., 2002) presupposed minimizing pairwise squared distances of points in each of k clusters where k is the user-specified parameter. In the standard implementation, the algorithm starts at a random set of k cluster centers, computes the closest center to each data points, and recalculates the centers. This sequence of steps continues until convergence or for a fixed number of steps. The functionality of the Mean-Shift method (Comaniciu , Meer, 2002), for which we used the approach provided by B. Feldman as a MATLAB function, is roughly the same, however, with quite different priorities. At the beginning, all data points are unassigned.

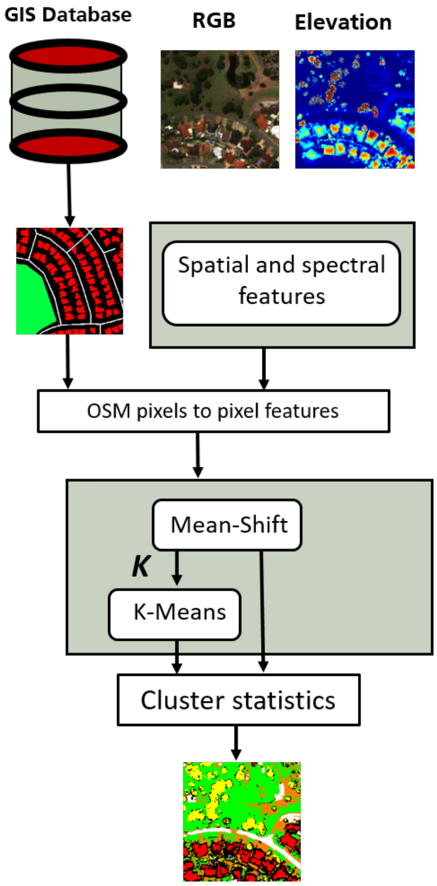

Figure 1: Process chain: Training data generation, fusion of GIS data, spectral and spatial features; clustering of pixels inside GIS areas; statistical analysis of cluster results.

An unassigned random point is selected and those data points within a certain range around it are assigned a new cluster. Its center is recalculated according to the new clusters' center of gravity. After this, the process is repeated until the cluster center does not move (convergence), whereby it must be checked if the current cluster must be merged with one of the previously retrieved. These steps are repeated until there are no unassigned clusters.

The main advantage of K-Means, namely its speed, is shadowcast by two disadvantages: the sensitivity against the initial choice of the cluster centers and of the unknown parameter k. The only parameter of Mean-Shift, the bandwidth, can be chosen more easily if some information about data accuracy (covariances) is available. The winner-takes-all mechanism is not advantageous in pathological situations (assume an 8-formed shape); however, in the majority of cases, its results are more stable. The only disadvantage is the running time. The authors try to accelerate this function by pre-computing neighboring structures using kd-trees has not been successful for big data because of memory requirements. In the following, we will show how cautious and meaningful handling of both tools allows compensating their shortcomings when it comes to unsupervised heterogeneous data classification.

3.2 Feature Extraction and Adaption for Clustering

Usually, land cover classification is based on the combination of features and a variety of features improves the results significantly. In the dataset considered in this work, a multispectral image was available from which we considered the four bands Red, Green, Blue and Near Infrared, which we will here denote as R, G, B and N respectively. From these features, other meaningful variables can be derived that are extensively used in remote sensing. These are NDVI (normalized difference vegetation index), NDWI (normalized difference water index), and NDSI (normalized difference soil index) (Wolf, 2012):

$$\text{NDVI} = \frac{N-R}{N+R}, \ \text{NDWI} = \frac{G-N}{G+N}, \ \text{NDSI} = \frac{G-Y}{G+Y}, \quad (1)$$

whereby Y means yellow and is computed by transforming the RGB image into the CMYK (Cyan, Magenta, Yellow, Black) color space. Additionally, we will occasionally employ the hue channel H of the HSV (Hue, Saturation, Value) color representation, whose advantages have already been successfully demonstrated by (Iwaszczuk et al., 2018) even though value channel was replaced by elevation.

In case of available elevation information, for example, a LiDAR point cloud, it can be resampled into DSM (digital surface model) and from there, the DTM (digital terrain model) and NDSM (normalized digital surface model) are calculated using tools of (Bulatov et al., 2012). These are our main spectral 3D features contrarily to 2D ones mentioned in the previous paragraph.

Additionally to spectral features, many authors (Bulatov et al., 2019, Fröhlich et al., 2013) employ the so-called spatial variables, which exploit a pixel's local neighborhood. In 3D case, the planarity feature (P) is computed from the structure tensor in order to assess how well the 3D points surrounding may be approximated by a plane (West et al., 2004). An example for a texture-based 2D feature is the entropy (E), which we mention for the sake of completeness in our Table 1. It turned out that it only barely provides an improvement, so we do not use it in our further processing. Due to the fact that the value ranges of the derived features are not scale-invariant, some of these features have to be normalized. According to the normalization, we have to consider the importance of the included features, e.g. NDSM distinguish between high and low land cover areas.

Table 1: Applied features

	2D	3D
spectral	R, G, N, H, Y NDVI, NDSI, NDWI	NDSM
spatial	E	P

First, we supposed that the used features are normally distributed, the *Z-transform* was used for normalization but led to unsatisfactory results. Looking in more detail at the features showed that a normal distribution cannot be expected at all (see Figure 2). Independent features, captured from different sensors have to be normalized due to their value ranges and dispersion feature. In case of RGB imagery, the three channels can obtain values between 0 and 255. In our research, NDSM values reach from about 0 to 24 meters. Considering those conditions, for each feature quantiles Q_F with at most 95% cumulative probability are derived. Those features are scaled the equation below:

$$Features^i_{norm} = \frac{Features^i}{Q^i_F}. \quad (2)$$

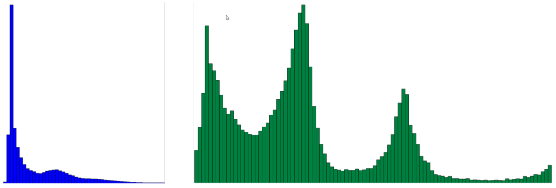

Figure 2: Histograms of NDSM (left) and Hue of HSV colorspace (right).

Finally, we use following features, *NDSM*, *NDVI*, *NDSI*, *NDWI*, *P*, *E*, *H*, *R*, *G* and *Y*, which are selectively included during the clustering process concerning the different classes road, soil, high vegetation (trees), and low vegetation (grass), building and water bodies.

3.3 Road and Building Extraction

Independently of the road's texture, luckily many roads with different appearances are contained by freely available GIS data (OSM) and can provide training examples for upcoming classification. Due to the fact that roads are stored as polygonal chains in OSM data, a fixed width value, depending on the image resolution, is attached to them to produce a raster image. Based on the width and image dimensions, the GIS road mask (background pixels = 0, road-pixels = 1) is generated. This process is performed using the *Bresenham* algorithm (Bresenham, 1965). For buildings, this task is more straightforward since fast algorithms exist allowing an assessment whether a point lies within a closed polygon. Empirically, we found out that for road extraction, most important spectral features are NDVI since it allows distinguishing between low vegetation and road, NDSI (especially for regions with forest/gravel roads), and NDSM. For the discrimination between low vegetation and roads with respect to the surface roughness, the planarity feature P is used. Based on the rasterized road mask, the features

$$F_{\text{road}} = (\text{NDVI}, \text{NDSI}, \text{NDWI}, Y, \text{NDSM}, P), \tag{3}$$

together with the aforementioned weights, are combined to be the main input for the Mean-Shift algorithm. First considerations how to determine the bandwidth h, was to use *Scott's normal reference rule* (Scott, 2015)

$$h = \frac{3.5 \cdot \sigma}{\sqrt[3]{N}}, \tag{4}$$

whereby σ is the standard deviation of used data and N is the number of data points. With this approach, the bandwidth h was approximately 0.1 (0.0969) and at first, it was used as a reference value. During further studies, concerning the value 3.5 and the cardinality of used root (equation 4). In (Martinez , Martinez, 2015), a modification of *Scott's normal reference rule* is described, changing the roots' cardinality (equation 5).

$$h = \frac{2.15 \cdot \sigma}{\sqrt[5]{N}}, \tag{5}$$

Mentioning that the features may be characterized by a skewed distribution, the bandwidth h is multiplied by the so called skewness factor (Martinez , Martinez, 2015). Based on our empirical studies, varying the number of features and, depending on the OSM mask, the number of pixels, the use of *Scott's normal reference rule* could be confirmed.

In the following step, all MS-clusters are analyzed with respect to their cardinalities and smaller ones are suppressed. Since outliers may occur, the cardinalities n_i of the resulting clusters are compared to the total number N of data points. If the ratio n_i/N is lower than a minimum $\gamma_{\min}$, these cluster pixels are ignored and this cluster is removed. The resulting number of Mean-Shift (MS) clusters together with the corresponding feature (equal to Mean-Shift data) determines the input parameters for the K-Means (KM) algorithm. Clustering results, the centers of gravity $\bar{F}_j^{\text{MS}}$ and $\bar{F}_k^{\text{KM}}$ (j, k: cluster indices) are analyzed whether the clusters of the MS and KM can be fused. The distances between $\bar{F}^{\text{MS}}$ and $\bar{F}^{\text{KM}}$ are determined. The KM-clusters, whose centers are below the specified distance β_{MS} to MS-centers, are fused. The cluster index of fitting KM indices C_k^{KM} are updated to the MS indices C_j^{MS}. This fusion process may lead to small shifts towards MS- and KM-clusters; therefore, the standard deviations σ_{feat} for the new clusters $C_j^{\text{MS+KM}}$ have to be updated. Based on the number of $C_j^{\text{MS+KM}}$ and corresponding σ_{feat}, constraints for road mask derivation are performed.

The rasterized OSM building masks can be processed in a completely analogous way, however, with a slightly different feature set (equation (6)).

$$F_{\text{building}} = (\text{NDVI}, \text{NDSI}, \text{NDSM}, P). \tag{6}$$

3.4 Soil and Vegetation Extraction

Concerning the geographic surroundings, soil areas may constitute both natural areas and access roads or gateways to buildings. Nevertheless, they shall be distinguished from roads, impervious surfaces, and low vegetation. Because soil areas are not a part of GIS data, natural areas denoted in OSM data as *landuse.shp* are inspected separately. Beside mentioned soil areas, high and low vegetation will also be present in these regions. Thus, during the training data extraction process for soil, vegetation types mentioned above may be extracted as well. The training masks extraction process can be divided into three steps:

1. If landuse data is available in the OSM data, a binary landuse map is generated. For all pixels labeled by 1, following features are chosen for clustering:

$$F_{\text{soil}} = (\text{NDVI}, \text{NDSI}, \text{NDSM}, P, Y). \tag{7}$$

2. This feature data is transferred to Mean-Shift and K-Means algorithm. After deriving distances between cluster centers of gravity $\bar{F}_j^{\text{MS}}$ and $\bar{F}_k^{\text{KM}}$, we perform fusion of clusters analogous to previous section and the updated cluster centers respective statistics are recomputed (see Figure 1)

3. According to updated cluster centers and their corresponding standard deviations for applied features, soil masks for each cluster centers are derived.

In Figure 3, orthophoto (left), the clustering result of Mean-Shift and K-Means (middle) and a detailed view of the soil masks is displayed (right). The non-dark-blue pixels of the middle image are those using for the K-Means clustering process. The green-dotted line in the orthophoto defines the outline of the natural area, the red outline reflects the boundary of the detailed view showing the final result of the superposed soil masks. The different colors are randomly set except the white color which denotes *no soil class*. Based on the clusters C_i, the soil masks SL_i are generated using centers $\bar{F}_i$ and standard deviations σ_i of features of data points in clusters as shown in Equation (8) below.

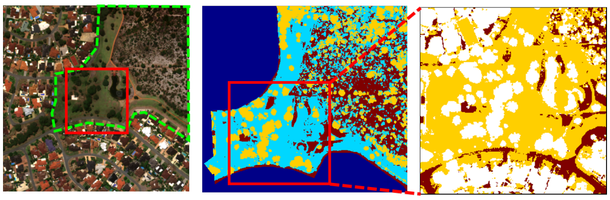

Figure 3: Detailed view: orthophoto; middle: clustering result; right: soil masks

$$\mathrm{SL}_i(x,y) = \begin{cases} 1 & \begin{array}{ll} \text{if} & \mathrm{NDSM}(x,y) < \overline{\mathrm{NDSM}}_i + \sigma_i^{\mathrm{NDSM}} \\ \text{and} & P(x,y) \geq \bar{P}_i - \sigma_i^{P} \\ \text{and} & \mathrm{NDSI}(x,y) \geq \overline{\mathrm{NDSI}}_i - \sigma_i^{\mathrm{NDSI}} \\ \text{and} & \mathrm{NDVI}(x,y) < \overline{\mathrm{NDVI}}_i + \sigma_i^{\mathrm{NDVI}} \\ \text{and} & Y(x,y) < \bar{Y}_i + \sigma_i^{Y}, \end{array} \\ 0 & \quad\;\; \text{otherwise.} \end{cases} \tag{8}$$

In case of vegetation, all pixels lying inside the landuse area are taken into account despite the fact that here the yellow channel was ignored and not used for Mean-Shift and K-Means. After the clustering, centers $\bar{F}_i$ ($\overline{\mathrm{NDSM}}_i$, $\bar{P}_i$, $\overline{\mathrm{NDSI}}_i$, $\overline{\mathrm{NDVI}}_i$, $\bar{Y}_i$) and the standard deviations σ_i^F of C_i are computed. To distinct between high vegetation (HV) and low vegetation (LV), the crucial features are NDSM and P.

$$\mathrm{HV}_i(x,y) = \begin{cases} 1 & \begin{array}{ll} \text{if} & \mathrm{NDSM}(x,y) \geq \overline{\mathrm{NDSM}}_i - \sigma_i^{\mathrm{NDSM}} \\ \text{and} & P(x,y) < \bar{P}_i + \sigma_i^{P}. \end{array} \\ 0 & \quad\;\; \text{otherwise,} \end{cases} \tag{9}$$

$$\mathrm{LV}_i(x,y) = \begin{cases} 1 & \begin{array}{ll} \text{if} & \mathrm{NDSM}(x,y) < \overline{\mathrm{NDSM}}_i - \sigma_i^{\mathrm{NDSM}} \\ \text{and} & P(x,y) \geq \bar{P}_i + \sigma_i^{P}, \end{array} \\ 0 & \quad\;\; \text{otherwise.} \end{cases} \tag{10}$$

Similarly to Equation (8), the NDVI values have to be greater than the corresponding cluster center values. All single masks belonging to SL_i, HV_i and LV_i are superposed.

3.5 Extraction of Water Bodies

Additionally to the aforementioned classes soil, building, grass, tree, and road, we also decided to extract the water surfaces, for which there existed no OSM data, while in the future it will be interesting to see whether our statistical approach may work in this case, too. Bearing in mind that NDWI values for water bodies must be relatively high, we visually estimated the portion of water surfaces in the data and chose the NDWI threshold corresponding to its quantile value. Since in the dataset considered in by our work, only a small lake and several pools make up water surfaces, this value was quite high (0.99). The connected components were finally assessed using their NDSM measure.

4. RESULTS

The aerial image combined with elevation data, we used for our research was captured over the City of Melville, which is a suburb of Perth, Australia. The spatial resolution of the image is 0.5 m and the image shows a residential area in the south and a park located in the north. Roads or forest ways exhibit on the one side asphalt or concrete surface while the roads in the inner of the parks resemble bare soil. Similar textures can also be seen in parking areas near to buildings. Considering the building roofs in the residential area, the roofs exhibit totally different roof colors and shapes. We used OSM data to localize those image pixels which belong to the corresponding OSM classes and then applied our approaches for unsupervised classification (sections 3.3 and 3.4).

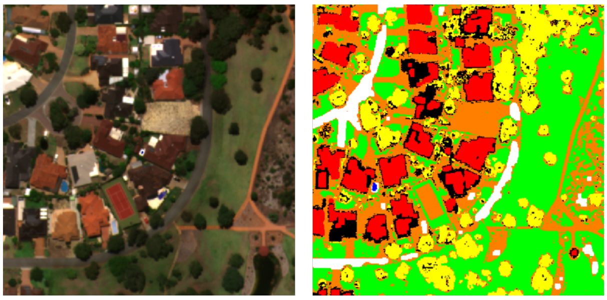

Figure 4: Detailed views: left: orthophoto; right: details of corresponding masks (white: road, red: building, green: gras, yellow: trees, brown: soil)

Next, we need ground truth. Since no largescale ground measurements were performed we relied on the approach (Bulatov et al., 2019), which was applied by (Kottler et al., 2019) to the Melville data and provided to the authors. Despite there were some small errors, it is good enough for a qualitative performance assessment of the proposed algorithm.

In Figure 5, two consciously selected areas superimposed in our result and ground truth are displayed. The differences are coded according to color: red = true positives (confirmed results), blue = false positives (or incorrectly assigned pixels), orange = false negatives (missed pixels), and white = false negatives. At a first glance, the major silhouettes are detected and especially orange parts, can be easily improved using morphological dilations. This is true for buildings and trees. At the same time, incorrect results mostly are typical for classes soil and grass and can be partly corrected by morphological erosion. In Figure 5 (results: 1st and 2nd row), one can easily identify a large overlap in the road mask. Nevertheless some road areas concerning e.g. sideways to buildings were not classified correctly. Contrary to Figure 5 (results: 3rd and 4th row), less roads are confirmed which can be related to various road textures or neighboring grass areas. Most of the buildings could be confirmed but also here there are some shortcomings caused by complex roofs and following low planarity values (see Figure 5, 2nd and 4th row, 3rd image), outer roof regions were not confirmed. Large grass areas (see Figure 5) are classified correctly. However, in case of a small lake which is present in the scenery (Figure 5, 3rd row, left corner of the orthophoto), the lake was misclassified as grass texture (Figure 5, 3rd row, 3rd image). Probably, it is because of abundant vegetation on the water surface. Trees could be detected correctly. In some cases, the center of the treetop exhibit large planarity values which led to holes in the tree crown (Figure 5,1st and 3rd row, right). For both areas, there is a large overlap concerning soil areas (see Figure 5, 2nd and 4th row right). Unfortunately, areas around buildings are incorrectly declared as soil. Finally, as we mentioned before, some earthy surfaces are declared as roads in the OSM data. Example are footpaths in parks and, in particular, the road on the margin of the park, as depicted in Fig. 5. In the ground truth, this belongs almost completely to soil area while after applying our approach, it is split between soil and road.

Melville

	R	*B*	*G*	*T*	*S*	*W*	*UA*
R	52104	286	13680	0	31971	588	52.8 %
B	87	96872	1048	663	37190	326	71.1 %
G	18	25	121119	1283	7787	0	93 %
T	7	72	17899	83246	48075	7	55.8 %
S	19666	1445	48520	0	140833	6	66.9 %
W	417	69	2008	0	537	2125	41.2 %
PA	72.1 %	98.1 %	59.3 %	97.7 %	52.9 %	69.6 %	68.9 %

Table 2: Confusion matrix for Melville (R: road, B: building, G: grass, T: tree, S: soil, W: water, UA: user accuracy, PA: producer accuracy)

Based on the result displayed in the confusion matrix shown in Table 2, we achieve an overall accuracy of 68.9 %. To complete the confusion matrix, producer accuracy (PA) and user accuracy (UA) are added. While evaluating the confusion matrix, it has to be noticed that the unlabeled pixels of our result were not included into the calculations of the confusion matrix. Referring to Table 2, it can be noticed that buildings could be separated from trees, but looking more precisely to the road class, one can easily see that a lot of soil-labeled pixels were misclassified as road class. Most remarkable miss-classifications occur between soil, grass and road.

5. CONCLUSION AND DISCUSSION

In our work, we presented a semi-automatic approach for training data acquisition using OSM data and actual airborne image and elevation data. We used combined image and elevation data to derive spectral and spatial features. For clustering, we used Mean-Shift and K-Means, whereby the result obtained with Mean-Shift determined the number of classes for K-Means and for the bandwidth parameter of Mean Shift, best results were obtained with the equation of Scott. The fusion result based on both clustering algorithms was analyzed.

For the mask generation concerning road, building, low-and high vegetation and soil, statistical analysis of the pixels belonging to the clusters C_i was carried out. After this, generation of the class masks took place. Nevertheless there still remain unlabeled pixels. Due to the fact that road textures are similar to soil and even dry grass textures, some pixels inside the natural area were misclassified and denoted as road pixels. Comparable results could be observed in the residential area. Here, more effort has to be put into integration of further features enabling a refined and better distinction between classes with low NDSM values and quite similar natural shades.

Despite the fact that the Mean-shift algorithm only needs the feature vectors itself and the bandwidth, more research must be done about feature normalization and bandwidth determination.

ACKNOWLEDGEMENTS

We thank to City of Melville and EagleView for providing the data of their municipality and to Rebecca Ilehag for data preprocessing.

REFERENCES

Audebert, N., Le Saux, B., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. *Asian Conferenceon Computer Vision*, Springer, 180–196.

Bresenham, J.E., 1965. Algorithm for computer control of a digital plotter. *IBM Systems journal*, 4, 25–30.

Bruzzone, L., Carlin, L., 2006. A multilevel context-based system for classification of very high spatial resolution images. *IEEE transactions on Geoscience and Remote Sensing*, 44, 587–2600.

Bulatov, D., Häufel, G., Lucks, L., Pohl, M., 2019. Land Cover Classification in Combined Elevation and Optical Images Supported by OSM Data, Mixed-level Features, and Non-local Optimization Algorithms. *Photogrammetric Engineering & Remote Sensing*, 85, 179–195.

Bulatov, D., Wernerus, P., Gross, H., 2012. On applications of sequential multi-view dense reconstruction from aerial images. *ICPRAM*, (2), 275–280.

Cheriyadat, A.M., 2014. Unsupervised feature learning foraerial scene classification. *IEEE Transactions on Geoscienceand Remote Sensing*, 52, 439–451.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions onPattern Analysis & Machine Intelligence*, 603–619.

Deng, Y., Wu, C., Li, M., Chen, R., 2015. RNDSI: A ratio normalized difference soil index for remote sensing of urban/suburban environments. *International Journalof Applied Earth Observation and Geoinformation*, 39, 40–48.

Fröhlich, B., Bach, E., Walde, I., Hese, S., Schmullius, C., Denzler, J., 2013. Land cover classification of satellite images using contextual information. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3, W1.

Geofabrik, 2017.OpenStreetMap-Shapefiles.www.geofabrik.de /en/data/shapefiles.html. Accessedat :2018-11-08.

Grippa, T., Georganos, S., Zarougui, S., Bognounou, P., Diboulo, E., Forget, Y., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., 2018. Mappingurban land use at street block level using open-street map, remote sensing data, and spatial metrics. *ISPRS International Journal of Geo-Information*, 7, 246.

Häufel, Gisela, Bulatov, Dimitri, Pohl, Melanie, Lucks, Lukas, 2018. Generation of training examples using osm data applied for remote sensed landcover classification. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 7263–7266.

Helmholz, Petra, Rottensteiner, Franz, Heipke, Christian, 2014. Semi-automatic verification of cropland and grassland using very high resolution mono-temporal satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 97, 204–218.

Iwaszczuk, D., Koppanyi, Z., Gard, N.A., Zha, B., Toth, C., Yilmaz, A., 2018.Semantic Labeling of Structural Elements in Buildings by Fusing RGB and Depth Images in AN Encoder-Decoder CNN Framework. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning A erial Image Segmentation from Online Maps. *IEEE Transactions on Geoscience and Remote Sensing*.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine In-telligence*, 881–892.

Kottler, B., Burkard, E., Bulatov, D., Harakè, L., 2019. Physically-based thermal simulation of large scenes for infrared imaging. *GRAPP 2019 - Proceedings of the International Conference on Computer Graphics Theory and Applications, Angers, France, May 17-21, 2010*, 53–64.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097 –1105.

Li, Y., Møgelmose, A., Trivedi, M. M, 2016. Pushing the "Speed Limit": high-accuracy US traffic sign recognition with convolutional neural networks. *IEEE Transactions on Intelligent Vehicles*, 1, 167–176.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lopes, P., Fonte, C., See, L., Bechtel, B., 2017. Using OpenStreetMap data to assist in the creation of LCZ maps. *Urban Remote Sensing Event (JURSE), 2017 Joint*, IEEE, 1–4.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. High-Resolution Semantic Labeling with Convolutional Neural Networks.. *arXiv preprint arXiv:1611.01962*

Marcos, D., Volpi, M., Kellenberger, B., Tuia, D., 2018. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145, 96–107.

Martinez, W.L., Martinez, A.R., 2015. *Computational statistics handbook with MATLAB*. Chapman and Hall/CRC.

Nayak, S., 2019. Number of Parameters and Tensor Sizes in a Convolutional Neural Network (CNN). https: // www. learnopencv.com/number-of-parameters\discretionary{-}{}{} and-tensor-sizes-in\discretionary{-}{}{}convolutional-neural -network/. [Online;accessed12-March-2019].

Qiu, C., Schmitt, M., Mou, L., Ghamisi, P., Zhu, X.X., 2018. Feature importance analysis for local climate zone classification using a residual convolutional neuralnetwork with multi-source datasets. *Remote Sensing*, 10, 1572.

Schilling, H., Bulatov, D., Niessner, R., Middelmann, W., Soergel, U., 2018. Detection of vehicles in multisensor datavia multibranch convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1–18.

Scott, D.W, 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Wassenberg, J., Middelmann, W., Sanders, P., 2009. An efficient parallel algorithm for graph-based image segmentation. *International Conference on Computer Analysis of Images and Patterns*, Springer, 1003–1010.

West, K.F., Webb, B.N., Lersch, J.R., Pothier, S., Triscari, J.M., Iverson, A.E., 2004. Context-driven automated target detection in 3D data. *Automatic Target Recognition XIV*, 5426, International Society for Optics and Photonics, 133–144.

Wolf, A.F, 2012. Using worldview-2 vis-nir multispectral imagery to support land mapping and feature extraction using normalized difference index ratios. *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII*, 8390, International Society for Optics and Photonics, 83900N.

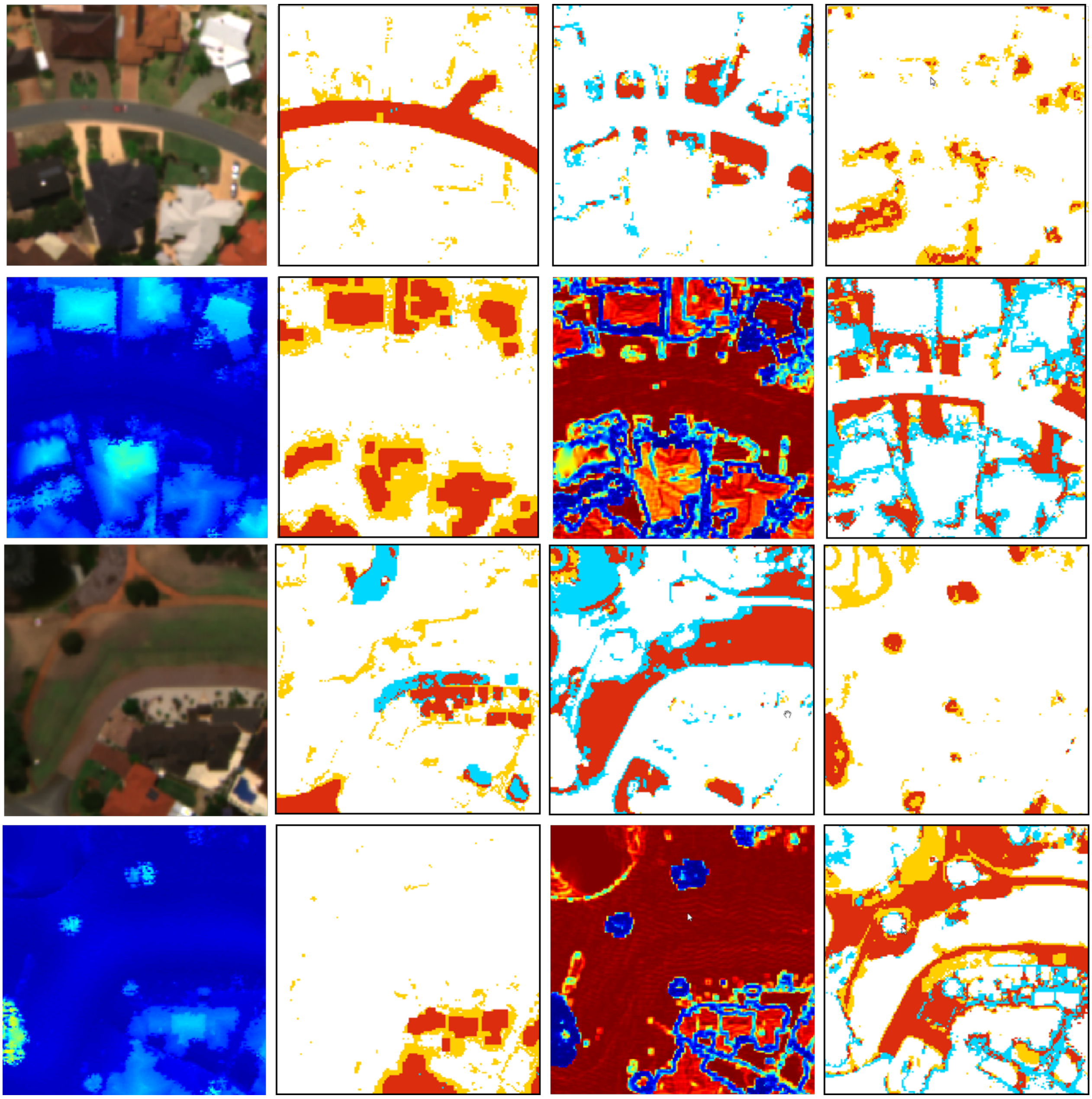

Figure 5: Results of area1 (1st and 2nd row) and area2 (3rd and 4th row): top row, left: orthophoto; middle to right: evaluation results for classes road, grass, and trees; bottom row, left to right: NDSM, evaluation result for the class building, planarity, evaluation result for the class soil.