FUSION OF LIDAR, HYPERSPECTRAL AND RGB DATA FOR URBAN LAND USE AND LAND COVER CLASSIFICATION

Sergey Sukhanov, Dmitrii Budylskii, Ivan Tankoyeu, Roel Heremans, Christian Debes

AGT International, Darmstadt, Germany
E-mail: {ssukhanov, dbudylskii, itankoyeu, rheremans, cdebes}@agtinternatonal.com

ABSTRACT

In this paper, we present an ensemble-based classification approach for urban land use and land cover classification based on multispectral LiDAR, hyperspectral and very high resolution RGB data. The approach has been evaluated on the data set provided for the IEEE GRSS 2018 Data Fusion Contest organized by the GRSS IADF technical committee and has been proven to have a high operational performance, being able to distinguish between different grass-, building- and street-types among other classes like water, railways and parking lots as well as other non-typical classes like cars, trains, stadium seats, etc.

Index Terms— multispectral LiDAR, hyperspectral imaging, very high-resolution RGB, land use classification

1. INTRODUCTION

Remote sensing the earth based on a multitude of different sensor technologies provides the means of measuring the earth in all its different facets going from visual over spectral and temporal characteristics to topological information like height. Through the past several years, the spatial resolution of all these sensors keeps improving dramatically providing an opportunity for precise land use and land cover classification of urban areas. As confirmation to that, one can observe that the ground sampling distance (GSD) of the hyperspectral image provided at IEEE GRSS 2013 Data Fusion Contests increased from 2.5 m in 2013 [1] to 1 m in 2018.

Despite these sensor-level improvements, multisensor-based classification is still a very challenging task in which data fusion approaches take a prominent role. Recently, traditional classification approaches such as SVM, Random Forests and different ensemble methods are being replaced by more sophisticated architectures of Neural Networks allowing to consider various aspects of multi modal data. This, however, often imposes limitations on the amount of data required to train such systems while providing significant potential of increasing classification performance.

In this paper, we propose an ensemble-based approach for urban land use and land cover classification that includes traditional classifiers as well as several advanced architectures of Neural Networks. The ensemble is based on Gradient Boosting Machines, a set of weak Random Forest classifiers and Convolutional Neural Networks. Various post-processing techniques, like mathematical morphology operators, region based property operators and Markov random fields, are applied in order to further improve the classification result.

The paper is structured as follows. In Section 2, we detail the source data and present the proposed framework in Section 3. We evaluate and discuss its performance in Section 4 and then conclude in Section 5.

2. DATASET

In this paper, we consider the data set provided by IEEE GRSS 2018 Data Fusion Contest technical committee [2] that contains heterogeneous types of remote sensing data covering the University of Houston campus and its surrounding areas. It includes RGB imagery, raw and processed Multispectral LiDAR data (acquired at three different wavelengths: 1550 nm, 1064 nm and 532 nm) and Hyperspectral imagery (HSI) covering a range of 380-1050 nm with 48 bands. RGB data is provided with 5 cm GSD and organized into several separate tiles. Multispectral LiDAR data is available as raw Multispectral LiDAR point cloud (LPC) data with 1 cm GSD as well as intensity rasters per channel: DSM, two types of digital elevation model (DEM) with different interpolation approaches and one hybrid DEM. All of the rasters are provided with 0.5 m GSD. HSI data is provided with 1 m GSD.

For training, a set of annotated pixel labels are provided representing distributed segments over a raster of 4768×1202 pixels at 0.5 m GSD resulting in $N = 2,018,910$ labeled pixels. In total, they span $M = 20$ unique urban land use and land cover classes (e.g. residential buildings, roads, artificial turfs, cars, etc.) demonstrating very unequal distribution of classes: around 15% of annotated pixels belong to the non-residential buildings, while less than 2% among the all annotated pixels stand for artificial turf, water, cross-walks and unpaved parking lots. The test set covers a region of 8344×2404 pixels at 0.5 m GSD having $341,729$ of them labeled.

3. PROPOSED CLASSIFICATION FRAMEWORK

The proposed multi-sensor urban land use and land cover classification approach is based on an ensemble of several

classifier instances as well as on a set of preprocessing techniques, feature extractions and postprocessing strategies that we detail in the sequel.

3.1. Preprocessing and feature extraction

LPC data was downsampled to 0.5 m GSD by applying median operator to altitude, intensity and number of returns. In addition, a distribution of coarse classification labels was calculated and later normalized by the number of aggregated points. Since after downsampling of LPC data there were still missing values a cubic and nearest neighbor interpolation were applied to fill them inside the resulting image and on the edges, respectively. The raw HSI data was upsampled to 0.5 m GSD by performing bi-cubic interpolation. The intensity values of raw RGB image were standardized per channel by subtracting mean and dividing by variance of red, green and blue channels of training data respectively.

Based on the LiDAR intensity channels the following set of features was extracted [3]:

- Pseudo NDVI (normalized difference vegetation index)
- Pseudo NDBHI (normalized difference built-up index)
- Combinations of intensity ratio for the three channels
- Brightness

In addition, we calculated ground height difference as a difference between raster values of DSM and DEM. Furthermore, we created several features out of the raster-based calculation by computing the difference of individual DEMs and DSM.

Finally, Morphological Texture Contrast (MTC) features [4] were applied to the intensity signal of every channel. MTC features were shown to have unique properties useful for discrimination of texture regions, e.g. forests or urban areas in aerial images, from background. Particularly, they are robust to spurious isolated edges or blobs and allow accurate segmentation at texture borders.

3.2. Cross-validation

To evaluate the models performance and estimate the weights of ensemble members, three cross-validation (CV) strategies were utilized. The first one was pixel-wise stratified 5-fold CV without shuffling that allowed to treat different parts of the labeled image separately making more difference between training and validation data distributions. The second CV strategy was pixel-wise stratified 5-fold CV with shuffling which caused validation data distribution to be more similar to the training data. The third strategy was based on superpixel (rectangular patch of pixels) pseudo-stratified 5-fold CV with shuffling. This strategy was particularly useful for context-based models since apart from the labeled pixel, a surrounding context of neighborhood was used for training. Due to the fact that superpixels are non-overlapping elements we padded them with reflection to allow edge-pixels having context. The label of each superpixel was assigned based on the dominant class label presented within this superpixel. Later, based on this labeling, stratified splitting was performed.

This approach can be considered as a different view of the dataset – multiple independent maps (i.e. superpixels) splitted into train / validation sets instead of one solid map representing all the region.

Since only superpixel-based approach could be properly used for context-based models, at the end it was the major approach for comparing all models performances, including pixel-wise based classifiers.

By averaging confusion matrices over different CV folds we estimated the performance of each model on every class and used this as weighting criteria for subsequent model ensemble.

3.3. Classification

The overall classifier is based on an ensemble of three different classifier instances: set of Random Forest classifiers, Gradient Boosting Machines and a set of Convolutional Neural Networks. In the following, we specify training process of each classifier instance and the way we combined them.

3.3.1. Set of Random Forest (RF) Classifiers

RF is a powerful classifier that builds and then averages over multiple decision trees to produce a classification output. Due to inherent bagging and random feature selection techniques the decision trees that form RF usually achieve high degree of diversity allowing RF to be susceptible to outliers and prevent overfitting. At the same time, RF (as well as the other discriminative classification methods) may suffer from class imbalance problems appearing when the class distribution in the training data is significantly skewed. This often results in an inability of a classifier to discover minority classes. To address these issues, motivated by the work in [5] we created an ensemble of T RF classifiers each trained on the reduced version of the original training dataset. Every subset is created by randomly undersampling the majority class to the level of a randomly chosen minority class and serves as training data for the RF classifier $f_t, t \in \{1, \ldots, T\}$. To aggregate the resulting classifiers we use dynamic classifier selection mechanism by partitioning the original data space into S regions. For each region $s \in \{1, \ldots, S\}$, the most competent RF classifier is picked according to its accuracy score on the validation set. To establish the regions of competence we cluster the original dataset disregarding class labels into S clusters using k-means. During the prediction phase, the distance between S cluster centroids and a data object is calculated to obtain the output of the corresponding RF classifier.

3.3.2. Gradient Boosting Machines (GBM)

GBM [6] is another ensemble-based classifier built of a number of sequentially created boosted decision trees by minimizing a custom cost function via a gradient decent algorithm. GBM naturally combines many of the advantages of boosting and tree-based algorithms making it powerfull and reliable classifier that through the last years was applied in many domains including remote sensing applications [7].

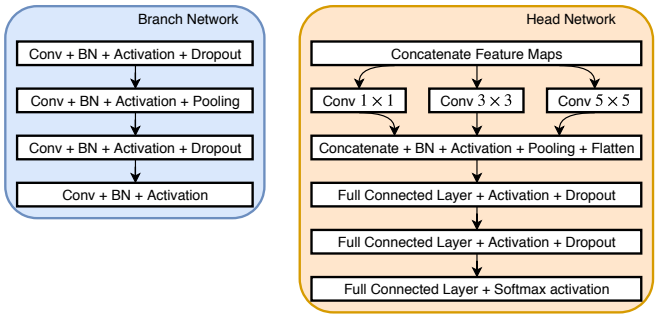

Fig. 1. Architecture of Branch Network (left) and Head Network (right)

3.3.3. Convolutional Neural Networks (CNNs)

Recently, CNNs have been shown to lead to outstanding classification performance when applied to remote sensing scenarios. In this work, we trained several CNNs each on a different data type separately. Generally, for every data type, the CNN input was represented by $x \in \mathbb{R}^{S \times S \times F}$, where $S = (2C + 1)R$, F is the number of features, C defines the radius (in 0.5 m^2 pixels) of the surrounding context and R is a data source scalar ($R_{HSI} = R_{LiDAR} = 1, R_{RGB} = 10$).

The branch CNN architecture for HSI- and LiDAR-based models is summarized as follows (see also Fig. 1):

- 64 convolution filters (CF) with kernel size $i \times i$ followed by Batch Normalization (BN) [8], activation $f_{activation}$ and Dropout [9] with probability $p_{dropout}$;
- 64 CF $(i \times i) \Rightarrow$ BN $\Rightarrow f_{activation} \Rightarrow$ Max Pooling;
- 128 CF $(i \times i) \Rightarrow$ BN $\Rightarrow f_{activation} \Rightarrow$ Dropout;
- 128 CF $(i \times i) \Rightarrow$ BN $\Rightarrow f_{activation}$.

Branch networks were created by varying filter size $i \in \{1, 3, 5, 7\}$ producing feature maps representing different local receptive fields. Filter sizes were chosen to represent local contexts of $1.5m \times 1.5m$ to $7.5m \times 7.5m$ with step of $2m$ which we supposed to be neither too small or too big. Head Network (Fig. 1) concatenates created feature maps, performs additional convolutional and dense transformations and outputs class probabilities. We used a categorical crossentropy loss function to train the network.

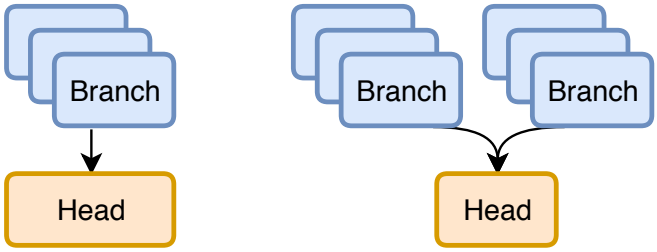

Fig. 2. (left) Lidar Network architecture draft, (right) HSI Network architecture draft

As LiDAR-based models we used Branch Networks (for each i) and single Head Network (Fig. 2, left) resulting into 2 CNNs with $C \in \{7, 11\}$ and Exponential Linear Unit (ELU) as $f_{activation}$. As HSI-based models we used two sets of Branch Networks: for $C = 3$ and for $C = 7$ followed by single Head Network (Fig. 2, right), having $f_{activation}(x) = ReLU(x) = max(0, x)$ resulting in three CNNs. As RGB-based model we used the Xception architecture [10] trained on patches with $C = 7$. For all networks $p_{dropout} = 0.2$.

3.3.4. Ensembling

The combination of the described models is done on the measurement level using weights for every class of each model obtained from the CV process. To obtain reliable confidence scores the posterior probabilities of all ensemble members were calibrated using isotonic regression. The overall ensemble consists of the following models described above: a set of RFs with dynamic selection, a GBM, two CNNs trained on LiDAR data, a CNN trained on HSI data and a CNN trained on RGB data.

3.4. Post processing

To improve the classification results of paved parking lots and artificial turf classes we apply two custom post-processing techniques. For paved parking lots we apply a $50 \times 50 \times 20$ window to every pixel of the posterior image calculating the amount of pixels where the confidence of class cars is the maximum. In case the amount of such pixels within the window is greater than a specified threshold $\varepsilon = 0.4 \times 50 \times 50$ then for the pixel under consideration the posterior for the class paved parking lot is set to 1 while for other classes to 0. Due to the fact that the artificial turf class was poorly represented in the training data many sport fields covered with artificial turf were wrongly classified as roads or paved parking lots. To account for that we extracted the binary maps of roads and paved parking lots, applied morphological opening and closing with square kernel of size 5×5 and calculated the area, extent and solidity for every disjoint segment on these maps to then set the posterior to 1 of the artificial turf class when these parameters exceed values of $10^4, 0.2$ and 0.6 respectively.

As a final step, to smooth the classification map, we modeled the label field as a Markov Random Field (MRF) where the Iterated Conditional Modes algorithm was applied to converge to a local solution [11].

4. EXPERIMENTAL RESULTS

To assess the performance of the proposed multi-sensor urban land use and land cover classification approach we evaluate it on the provided test set. For the set of RF classifiers and GBM only LiDAR features were used. We set $T = 100$, $S = 15$.

In Table 1 the overall accuracy and kappa measures on the test set are provided for the proposed ensemble methods as well as all ensemble members individually.

By analyzing the results we can conclude that, in general, the ensemble is able to accurately provide land use and land cover classification outperforming its individual members. Obviously, there are also several limitations that can be

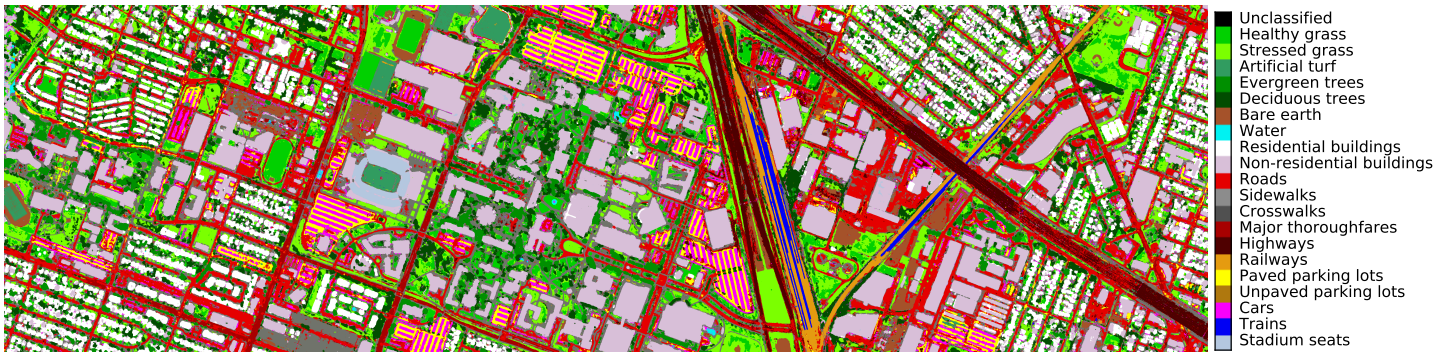

Fig. 3. Prediction map on the test set

	Overall accuracy	Kappa
Set of RF classifiers	69.37%	0.67
GBM	61.81%	0.59
CNN (LiDAR)	69.01%	0.67
CNN (HSI)	47.61%	0.44
CNN (RGB)	63.95%	0.62
Proposed ensemble approach	**79.79%**	**0.79**

Table 1. Experimental results

seen from the resulting prediction map (Fig. 3): Some classes are not well classified due to class imbalance (according to kappa measure) e.g. unpaved parking lots or crosswalks; major thoroughfares, roads and highways have mutual confusion since the actual class depends on the context only that in many cases is ambiguous.

5. CONCLUSION

In this paper, we proposed a multi-sensor urban land use and land cover classification approach based on an ensemble of multiple classifiers. We validated the proposed approach on the recent multi-sensor remote sensing dataset demonstrating high operational performance and generalization capabilities of the proposed method.

6. ACKNOWLEDGEMENT

The authors would like to thank the National Center for Airborne Laser Mapping and the Hyperspectral Image Analysis Laboratory at the University of Houston for acquiring and providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

7. REFERENCES

[1] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pizurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE JSTARS*, vol. 7, no. 6, pp. 2405–2418, 2014.

[2] "2018 IEEE GRSS Data Fusion Contest," `http://www.grss-ieee.org/community/technical-committees/data-fusion`.

[3] L. Matikainen, K. Karila, J. Hyypp, P. Litkey, E. Puttonen, and E. Ahokas, "Object-based analysis of multispectral airborne laser scanner data for land cover classification and map updating," *ISPRS J Photogramm Remote Sens*, vol. 128, pp. 298 – 313, 2017.

[4] I. Zingman, D. Saupe, and K. Lambers, "A morphological approach for distinguishing texture and individual features in images," *Pattern Recognition Letters*, vol. 47, pp. 129–138, 2014.

[5] S. Sukhanov, A. Merentitis, C. Debes, J. Hahn, and A. M. Zoubir, "Bootstrap-based SVM aggregation for class imbalance problems," in *Proceedings of EUSIPCO*, 2015, pp. 165–169.

[6] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[7] N. Yokoya, P. Ghamisi, J. Xia, S. Sukhanov, R. Heremans, I. Tankoyeu, B. Bechtel, B. L. Saux, G. Moser, and D. Tuia, "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS data fusion contest," *IEEE JSTARS*, 2018.

[8] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014.

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE CVPR*, 2017, pp. 1251–1258.

[11] S. Sukhanov, I. Tankoyeu, J. Louradour, R. Heremans, D. Trofimova, and C. Debes, "Multilevel ensembling for local climate zones classification," in *Proc. of the IEEE IGARSS*, 2017, pp. 1201–1204.