BENCHMARKING CLASSIFICATION OF EARTH-OBSERVATION DATA: FROM LEARNING EXPLICIT FEATURES TO CONVOLUTIONAL NETWORKS

Adrien Lagrange[1], *Bertrand Le Saux**[1], *Anne Beaupère*[1], *Alexandre Boulch*[1],
Adrien Chan-Hon-Tong[1], *Stéphane Herbin*[1], *Hicham Randrianarivo*[1], *Marin Ferecatu*[2]

[1] Onera – The French Aerospace Lab, F-91761 Palaiseau, France
[2]CNAM - Cedric, 292 rue St-Martin, 75141 Paris, France

ABSTRACT

In this paper, we address the task of semantic labeling of multisource earth-observation (EO) data. Precisely, we benchmark several concurrent methods of the last 15 years, from expert classifiers, spectral support-vector classification and high-level features to deep neural networks. We establish that (1) combining multisensor features is essential for retrieving some specific classes, (2) in the image domain, deep convolutional networks obtain significantly better overall performances and (3) transfer of learning from large generic-purpose image sets is highly effective to build EO data classifiers.

***Index Terms*—** Remote sensing, Image classification, Pattern analysis, Neural networks

1. INTRODUCTION: URBAN CLASSIFICATION

The study of urban centers using Earth-Observation (EO) data has a lot of potential users and applications, from urban management to flow monitoring, and in the meantime offers great challenges: numerous and diverse semantic classes, occultations or bizarre geometries due to the image-capture angle and the ortho-rectification. Semantic labeling consists in automatically building maps of geolocalized semantic classes. It evolved with both the resolution of data and the availability of labeled data. The contribution of resolution is straightforward: with more details, new potential semantic classes can be distinguished in the images: from roads and urban areas to buildings and trees. Then image description evolved from textures to complex features which allow object modelling [1, 2]. Meanwhile, labeled datasets allowed a rigorous validation of algorithms and the development of statistically-based methods for multi-class urban classification [3, 4]. More recently, very large training sets were used to train deep networks [5], for example based on convolutional networks [6].

Despite these impressive advances, semantic labeling still faces unsolved problems: which method is best suited for a given class ? Is it possible to build a classifier which is generic enough to handle a large variety of labels ? Semantic classes may have really diverse structures, from large, loose areas (i.e. vegetation areas) to rigid, structured objects (such as cars, street furniture, etc.). Actually, with the advent of very-high resolution (VHR) images, the latter becomes more and more frequent.

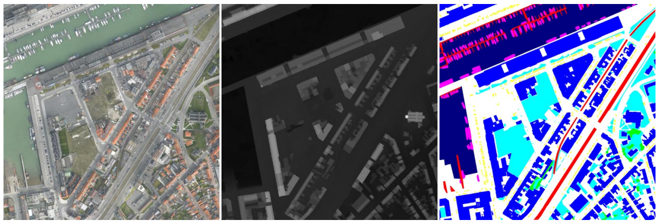

Fig. 1. *grss_dfc_2015* for data fusion: **(a)** orthophoto tile at 5cm-resolution along with **(b)** the corresponding DSM at 10cm-resolution (derived from LiDAR). **(c)** is the label map corresponding to the ground truth we built.

The VHR multi-sensor dataset provided in the framework of the IEEE GRSS Data Fusion Contest provides us with a large variety of semantic classes. In this study, we use it as the benchmark needed by the EO community for rigorously assessing and comparing the various approaches that coexist. For this purpose, we built up a ground truth with 8 classes (cf. Section 2) that we propose to make available. We implemented and tested various approaches ranging from expert and sensor-based baselines to powerful machine-learning approaches, aiming at both pixel-wise and object-wise classification (cf. Section 3). Their respective performances can be evaluated and compared (cf. Section 3.6) showing which ones are best suited for some specific applications and which ones are overall winners that could be used for generic purposes.

2. BENCHMARK

Dataset and ground truth. The IEEE GRSS DFC Zeebrugge dataset ([7], referenced in the following as: *grss_dfc_2015*)

*Corresponding author: bertrand.le_saux@onera.fr
All authors: <firstname>.<lastname>@<institution>.fr

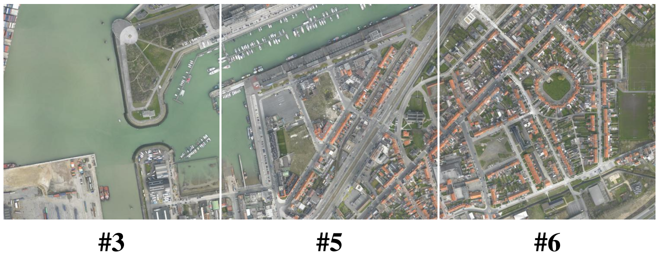

Fig. 2. Images from the *grss_dfc_2015* dataset. Middle image **#5** contains almost all the semantic classes of Table 1 and belongs to the training set. In the test set, image **#3** contains a harbour zone and image **#6** contains a large residential area.

contains 7 orthorectified tiles, with the following data:

- a 10000x10000 pixel-sized color orthophoto (RGB, 5cm-resolution).
- a max 5000x5000 pixel-sized Digital Surface Model (DSM) at 10cm-resolution.
- a LiDAR 3D-point cloud in XYZI format [containing X (latitude), Y (longitude), Z (elevation), and I (*LiDAR intensity*) information].

In addition, we manually built a ground truth (cf. Fig. 1) with semantic labels summarized in Table 1.

Evaluation. We perform cross-validation on the dataset to assess the various methods. We retain images $\{1, 5, 7\}$ for training and images $\{3, 6\}$ for testing. They are chosen for ensuring a good representation of all classes in both sets: for example, image 5 is the most representative of the semantic classes with harbour and residential areas, while image 3 contains a harbour zone and image 6 contains a large residential area (cf. Fig. 2).

Pixel-wise classification is evaluated using the confusion matrices for each image. We count (for each class or over the test set) the number of *true positive* pixels tp, the number of *false positives* fp, the number of *false negatives* (or miss) fn. We then derive different standard measures for each class: precision $(= tp/(tp + fp))$, recall $(= tp/(tp + fn))$, and the F1-score $(= 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall}))$. We also compute the overall accuracy $(= \frac{(tp+tn)}{(\text{total number of pixels})})$ and Cohen's *Kappa* $(= \frac{p(a)-p(e)}{1-p(e)}$, where $p(a)$ is the observed accuracy and $p(e)$ the expected accuracy), computed using the confusion matrix.

Table 1. Ground truth classes for semantic labeling (with class proportion over *grss_dfc_2015*).

⬜	🟦	🟦	🟩
Impervious surface	Building	Low vegetation	Tree
33.6 %	8.2 %	10.8 %	2.0 %
🟨	🟥	🟪	🟦
Car	Clutter	Boat	Water
0.5 %	7.8 %	0.7 %	28.7 %

3. ALGORITHMS AND BASELINES

We test several approaches for classification, from handcrafted heuristics to learning algorithms based on raw data or carefully-designed image descriptors.

3.1. Expert baselines

When possible, we build label-specific baselines. Most of them are single-channel filters on RGBd data. The *water* classifier checks if $d < 45.4m$. The *building* classifier checks if $d > 50.5m$. The *road* classifier (for impervious surfaces) look for gray pixels below a given depth: $max(R, G, B) - min(R, G, B) < 6$ and $d < 52m$. Assuming that most LiDAR systems for land observation have near-infrared (NIR) wavelengths, we projected the intensity from the LiDAR point-cloud to create pseudo-NIR images. We then computed the normalized difference vegetation index (NDVI) using $(NIR - R)/(NIR + R)$ and fixed the threshold at 0.6.

3.2. SVM on raw data

As a simple baseline, we train a Support-Vector Machine (SVM) on raw data. Various inputs are considered: RGB values of image pixels, RGBD values by adding the DSM, and RGBID values, where I is a pseudo-infrared derived from the Lidar intensities. One SVM is trained for each class in a one-vs.-all manner, using a Radial-Basis-Function (RBF) kernel with internal parameters optimized by grid-search. To prevent an explosion of the computational costs, classification is performed on the averaged value of superpixels computed using efficient graph-based segmentation [8].

3.3. SVM on complex features

We tested two approaches for high-level feature extraction.

1. In the spatial-spectral domain: patches (16×16 or 32×32) are extracted, indexed with Histograms of Oriented Gradients (HOGs) (implemented as in [9]) and given the dominant label. We then train several RBF-kernel SVMs in one-vs.-all set-ups with optimal parameters found by grid-search. At classification, we apply the classifier using a standard sliding window approach and smooth the resulting map.
2. Using multisource information: superpixels are computed on the image, then described by HSV-color histograms combined with the averaged value and averaged gradient of the DSM. The classifier is learned by a linear SVM.

3.4. Object-based detectors

We also tested 2 methods for object-oriented detection.

1. (Discriminatively-trained Model Mixtures) improve the work of [2] based on Discriminatively-trained Part Models

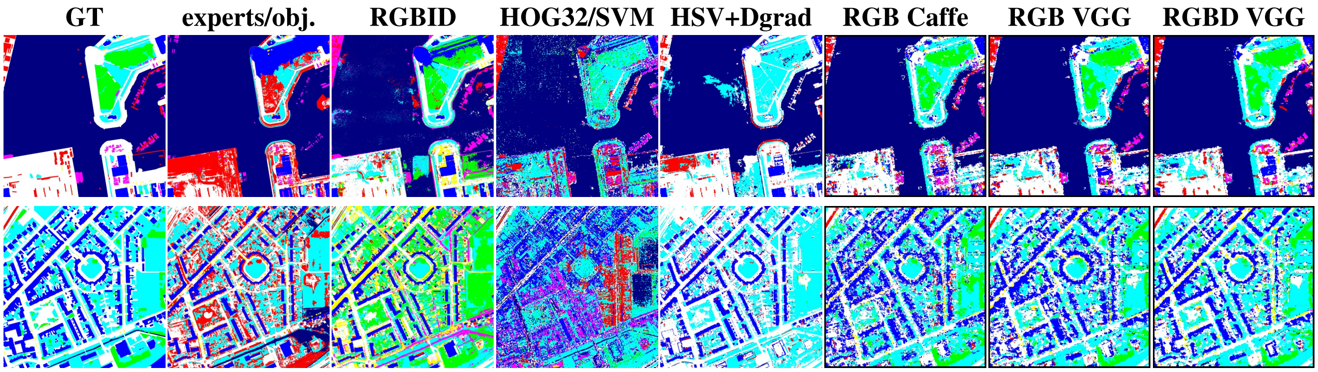

Fig. 3. Comparison of classification maps for image #3 (first row) and #6 (second row) of *grss_dfc_2015*, with respect to the ground truth (**GT**): (**experts/obj.**) experts and object classifiers combined on a single map, (**RGBID**) superpixels classified by SVM with RBF kernel, (**HOG32/SVM**) HOG features with RBF-SVM, (**HSV+Dgrad**) features computed on superpixels and classified with linear SVM, (**RGB Caffe**), (**RGB VGG**) and (**RGBD VGG**) CNN-features with linear SVM.

Table 2. Method comparison: F1 measures per class (best: 🟩, second: 🟩, third: 🟨), overall accuracy and Cohen's Kappa.

Algorithm	Imp. surf	Build.	Low veg.	Tree	Car	Clutter	Boat	Water	Overall acc. %	Cohen κ
Expert	58.97	63.87	74.55					92.39	∅	∅
RGB	53.89	53.53	50.32	32.97	24.02	13.75	12.12	98.52	60.77	0.52
RGBD	14.51	67.79	38.03	27.43	7.15	1.12	14.58	98.45	50.76	0.41
RGBID	60.86	69.01	57.12	38.12	11.59	20.49	15.04	94.42	63.83	0.56
HOG32/SVM	28.94	43.17	48.77	27.32	30.24	17.39	12.61	88.02	52.45	0.41
HOG16/SVM	39.52	38.45	35.65	29.99	21.93	16.13	13.52	80.02	49.4	0.36
HSV/SVM	71.60	46.97	68.38	0.12	0.00	13.71	0.00	92.14	70.16	0.60
HSV+Dgrad/SVM	73.30	70.85	68.75	0.17	0.00	17.11	0.00	92.37	73.60	0.65
SOM							51.45		∅	∅
DtMM					48.46				∅	∅
RGB OverFeat/SVM	55.86	63.34	59.48	64.44	36.03	28.31	41.51	92.07	67.97	0.59
RGB Caffe/SVM	62.32	62.66	63.23	60.84	31.34	32.49	46.57	95.61	71.06	0.63
RGB VGG/SVM	63.18	64.66	63.60	66.98	31.46	43.68	51.92	95.93	72.36	0.64
RGBD VGG/SVM	66.02	74.26	65.04	66.94	32.04	44.96	50.61	96.31	74.77	0.67
RGBD^{+} VGG/SVM	67.66	72.70	68.38	78.77	33.92	45.6	56.10	96.50	76.56	0.70

[9]. The model of an object category consists in a mixture of discriminative models trained on visually homogeneous data: object samples are clustered on the basis of the visual appearance and for each cluster a linear SVM is trained on HOGs computed on these samples.

2. The second object detector is based on Self-Organizing Maps (SOM): it learns an optimal color table for the images that can be used for segmenting the test images. Semantic labels are associated to SOM outputs on the training set, and derived from SOM classification maps.

3.5. Convolutional neural networks and SVM

In recent years, convolutional neural networks (CNN) have achieved the best performances on various benchmarks (e.g. everyday-image classification [10]). It has been experimented that the outputs of the intermediate layers of these deep networks could be efficiently used as features [11]. We use three different implementations of CNN trained on ImageNet [12] that we cut before the soft-max layer: VGG (5 convolutional-layer fast network [13]), OverFeat (the fast network of 6 conv. layers, no drop-out) [14] and Caffe (the network has 5 conv. layers [10]). We generate features on $231 * 231$ patches extracted from the training images by a sliding window approach (step of 32 pixels) and train a linear SVM with respect to our 8 classes. At testing, the same sliding window is used, and the resulting label is given to the central $32 * 32$ square of the patch. Moreover, we test the contribution of LiDAR. We apply VGG to the DSM, and trained a linear SVM over the concatenated output of RGB and depth networks. We use either the given DSM (RGBD) or a more precise DSM (RGBD$^+$) obtained by projecting height from the LiDAR point-cloud.

3.6. Results and analysis

In Fig. 3 we show the classification maps along with the ground truth, while Table 2 summarizes the performance measures for each class and overall. The use of superpixels introduce spatial constraints that are visually rewarding on classification maps, especially in dense urban environment (Fig. 3). Multisource information is a key to success: the two best approaches combine image and DSM. Working on images only, deep neural networks are solid candidates for building generic EO data classifiers. In Table 2, they often outperform the other baselines and get consistent results over the 8 labels, with a bonus for not-so-deep networks (5 layers) that are less specialized. The next question is how to use these neural networks in EO data context: either by transfering learning from large everyday-image datasets (as was performed here) or by retraining ? Finally, old recipes are still competitive on specific challenges (cf. Table 2): NIR information is crucial for vegetation, while depth and colorimetry are meaningful for buildings and water respectively. Moreover object-oriented methods perform well on the task they are designed for: While objects do not count for much in pixel proportion (cf. Table 1), they have a high interest in some critical applications.

4. CONCLUDING REMARKS

In this paper, we established a ground truth for semantic labeling associated with the *grss_dfc_2015* data that we propose to make available to the community. We tested various state-of-the-art approaches for urban classification on this challenging benchmark. The main outcomes are that: (1) multisource combination is highly relevant for some specific urban classes; (2) as a generic all-purpose classifier, deep convolutional networks obtain significantly good performances; and (3) transfer of learning from large generic-purpose image sets is highly effective to build EO data classifiers.

Aknowledgement

The authors would like to thank the Belgian Royal Military Academy for acquiring and providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

5. REFERENCES

[1] J. Leitloff, S. Hinz, and U. Stilla, "Vehicle detection in very high resolution satellite images of city areas," *IEEE Trans. on Geosci. Remote Sens.*, vol. 48, no. 7, 2011.

[2] H. Randrianarivo, B. Le Saux, and M. Ferecatu, "Man-made structure detection with deformable part-based models," in *Int. Geosci. Remote Sens. Symp.*, 2013.

[3] M. Fauvel, J. Chanussot, and J.A. Benediktsson, "Decision fusion for the classification of urban remote sensing images," *IEEE Trans. on Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2828–2838, 2006.

[4] D. Tuia, F. Pacifici, M. Kanevski, and W.J. Emery, "Classification of very high spatial resolution imagery using mathematical morphology and support vector machines," *IEEE Trans. on Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3866–3879, 2009.

[5] V. Mnih and G. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. of Eur. Conf. Comp. Vis.*, 2010.

[6] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction of hyperspectral images," in *Proc. of WHISPERS*, 2014.

[7] 2015 IEEE GRSS Data Fusion Contest, "Online: `http://www.grss-ieee.org/community/technical-committees/data-fusion`,".

[8] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J Comp. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Patt. An. Mach. Int.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[11] Jason Yosinski, Jeff Clune, Geoffrey Hinton, and Hod Lipson, "How transferable are features in deep neural networks?," in *Proc. of NIPS*, 2014, pp. 3320–3328.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. of Comp. Vis. and Patt. Rec.*, 2009.

[13] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Brit. Mach. Vis. Conf.*, 2014.

[14] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. on Learning Rep.*, 2014.