Structural classifiers for contextual semantic labeling of aerial images

Hicham Randrianarivo*, Bertrand Le Saux*, Nicolas Audebert*, Michel Crucianu† and Marin Ferecatu†
*Onera The French Aerospace Lab
F-91761 Palaiseau, France
† Conservatoire National des Arts et Métiers, Laboratoire CEDRIC
292 Rue St Martin FR-75141 Paris Cedex 03

***Abstract*—In this work we propose a novel approach to classify aerial images with structured predictors. They get high performances by encoding the contextual information that exists around a region of interest. More precisely, we learn a graph model that takes into account the structure existing between adjacent regions (or superpixels) belonging to various categories with a Structural Support Vector Machine. The whole image classification task is broken down in multiple local subtasks so that we can deal with large amounts of regions (more than 100k).**

I. Introduction and previous works

A. Introduction

This work focuses on semantic labeling of aerial images. Nowadays more and more images captured by aerial or satellite systems become available. For example the Sentinel satellites collect more than 2To of images a day. The spatial resolution of these images vary between 300 meters to few centimeters. Manually analyzing this amount of data would take a lot of time and manpower: this is why we want to automatically analyze these images. Our goal is to automatically classify all the pixels of an image into several predefined categories. The task of segmenting and labeling Very High Resolution (VHR) imagery is one of the most active research axes in the remote sensing community [1], [2], [3], [4], [5], [6] as it is essential for a wide range of activities like deforestation analysis or urban modeling. The standard approach consists in extracting a feature descriptor on several regions of the images then learning a model over the features in order to predict the category of the region of interest. Our approach is based on the idea that we can enrich the description of a region of interest with more information from the image. By incorporating in the model information about the neighborhood of the regions of interest we can improve the recognition abilitiy of a model and then improve its predictions. In this paper we show that even novel and state-of-the-art approaches like deep neural networks (AlexNet [7]) can be improved if we extract useful contextual information of the neighborhood of the region of interest. To this aim we learn the pattern which exists in the local neighborhood structure using a graph representation and a Structural Support Vector Machine (SSVM) framework.

B. Overview

Recently a new state of the art in semantic labeling has been established with deep neural networks. Such networks (for example AlexNet) are known to be very effective when it comes to extract a powerful representation of the content of an image, even when trained on general purpose datasets [5]. We use two kinds of setup based on AlexNet as our baseline algorithms. The first one compute features on small patches centered on the superpixels while the second one uses quite larger patches which include more context. In both cases feature descriptors are compute on the patches using AlexNet and used to train a multiclass Linear Support Vector Machine (SVM).

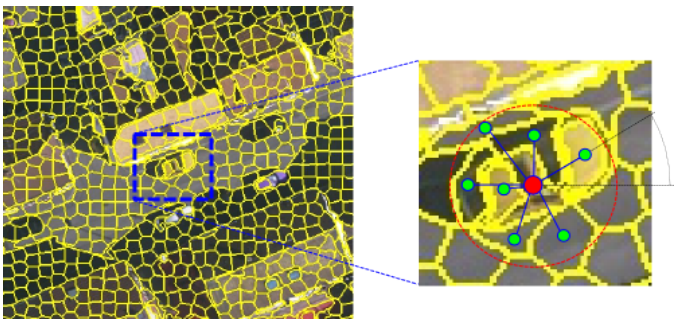


Fig. 1: Overview our method to model local relations between supepixels. For a superpixel of interest we extract a local graph of relations that will be used to make a prediction. The node of interest is the node in red and the neighbor nodes are in green. We model interactions between a node of interest and the neighbor nodes to improve the prediction.

Our approach builds on the small patch baseline but moreover encodes the context using higher order information. In an image, some categories of objects are more likely to appear next to other categories of objects (cars are more likely to appear on a road than on a tree). Our method proposes to model the structure that exists between differents regions in the image and to learn a contextual model that take into account the local interactions between the regions to predict a category. We model the local interactions between regions using a graph structure. The nodes of the graph are the feature vectors of the region of interest. Using a graph structure allows to add extra information on the edges which will be useful to model the interactions between the nodes. To this effect we define a contextual feature which captures the relative positions of the

doi: 10.2788/854791

regions that lie below a given radius. Locally for each region we regularize the local graph computed from single classifier using SSVM which is an extension of the popular SVM. The SSVM framework then predicts new classes for all regions of the graph.

II. Contextual modeling

A. Baseline classifier

Our baseline classifiers predict a label for each superpixel. The visual information associated with a superpixel is taken from the patch centered on it. It will be processed by the deep network. We use the following pipeline for semantic segmentation:

1) Divide the image into small regions of interest (superpixels) using the Simple Linear Iterative Clustering (SLIC) algorithm [8]
2) For a region of interested extract a patch of size $(N \times N), N \in \{32, 64\}$ centered on the superpixel.
3) Resize the patch to 228×228 and process them through AlexNet.

We use theses features in order to train a linear multiclass SVM. As a groundtruth we have label maps where all the pixels are assigned to a category. One issue with using superpixels as training samples is that superpixels can incorporate pixels from two differents categories. To deal with this problem we perform a majority vote according to the groundtruth to label a superpixel. The resulting semantic map is obtained thanks to the prediction of the SVM.

B. Structural context

Usually the region of interest (for instance a patch) we want to classify are very small parts of a semantic object. Looking only at patch level shadows or unusual appearance may lead to misclassifications. Introducing visual information from the neighboorhood allows to reduce uncertainty about the patch.

1) Graph model: The second contextual information we use in this work is the structure of the local interactions of regions of the image. The underlying assumption is that superpixels surrounding a region of interest have a structure that can be learnt. One natural datastructure to model structural relationship between several elements is a graph model. Using a graph to model interactions between superpixels allows us to incorporate extra contextual information in the graph. We construct the graphs that capture interactions between regions of an image in the following manner:

1) Divide the image into small regions using the SLIC algorithm [8]
2) For a region of interested extract a patch of size (32×32) centered on the superpixel.
3) Create the nodes of the graph using the features
4) Create edges between nodes where the distance between the centroids is less than a radius r
5) For each edges in the graph compute a feature of context composed of the distance ρ between the centroid and the angle θ which is the relative orientation between the superpixel centroids.

The result for each image is a set of nodes $\mathcal{V}$ and a set of undirected edges $\mathcal{E}$. A graph $G = (\mathcal{V}, \mathcal{E})$ is then associated with each image of the training set.

2) Structural model for context: Our model is composed by unary features (describing the nodes) and pairwise features (contextual feature between two nodes) which jointly describe interactions between input and output variables. For training, we use a set of N images associated with their label maps. From the images, we extract graph models as explained in section II-B1. We then extract a set of local relation graphs that we use as training samples $X = \{x^n\}_{n=1}^N$ with the corresponding groundtruth annotations $Y = \{y^n\}_{n=1}^N$ generated from the label maps. The specificity of the SSVM framework is that the output labels Y are structured, which means they are graph of classes and not single class values. A target y^n is a set a labels y_i where each label corresponds to a node x_i. The labeling of a region of interest is found by minimizing the following energy function:

$$E(x, y, ; w) = \sum_{i \in \mathcal{V}} \phi_i(y_i; w^\phi) + \sum_{(i,j) \in \mathcal{E}} \varphi_{i,j}(y_i, y_j; w^\varphi) \quad (1)$$

$$= \langle w, \psi(x, y) \rangle \quad (2)$$

With $\phi_i(y_i)$ as the unary term and $\varphi_{(i,j)}(y_i, y_j)$ as the pairwise term. The parameters to learn are w^ϕ and w^φ.

3) Max-margin structured learning: The max-margin structured learning framework optimizes discriminatively the weights of the energy function described in Equation (1). Learning the weight parameters of Equation (1) does not scale well because the computational cost is quadratic. The authors of [9] propose an efficient method to solve this issue using Block-Coordinate algorithm which allows to break down the optimization problem into simpler linear subproblems. The SSVM framework finds model weights that maximize the energy of any labeling y with respect to the one of the groundtruth y^n by the largest margin $\Delta(y, y^n)$

$$w* = \arg\min \frac{1}{2}||w||^2 + \frac{C}{N} \sum_{n=1}^{n} l(x^n, y^n, w) \quad (3)$$

with

$$l(x^n, y^n, w) = \max_{y \in \mathcal{Y}} \Delta(y^n, y) - g(x^n, y^n, w) + g(x^n, y, w) \quad (4)$$

$$g(x, y, w) = \langle w, \varphi(x, y) \rangle \quad (5)$$

where C is a penalized hyperparameter and $\Delta(y, y^n)$ is a loss function that measures the error of predicting y knowing the real configuration is y^n. Several loss functions have been defined for the max-margin structured problem. In this work we use the common Hamming loss which aims at penalizing wrong labeled nodes equally and is defined by $\Delta(y, y^n) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} [y_i \neq y_i^n]$

doi: 10.2788/854791

4) Predicting a label: For each superpixel of an unknown image we consider the local graph of neighbors inside a given radius. The SSVM framework predicts the labels of all the nodes in the graph of local interactions. One approach could be to keep only the label of the region of interest and using it to predict the label of the superpixels. In our approach we also predict the labels of surrounding superpixels, so for the whole image we get several predicted labels for each superpixel. We exploit this by setting up a voting procedure where for each superpixel we accumulate the votes of all the neighbors.

III. Experiments

A. Setup

The methods are tested on the ISPRS 2D Semantic Labeling Dataset [10]. We use part of the Vaihingen data, consisting of 16 *IR-R-G* orthoimages with pixel-level ground truth. We asset the quality of the classification with the ground truth.

We split this dataset as follows : tiles 1, 5, 7, 11, 17, 23,26, 28, 34 and 37 form the training set, while tiles 13, 21 and 30 form the validation set and tiles 3, 15 and 32 form the testing set. Note that the "clutter" class is not represented in the testing set. This is justified by the fact that the ISPRS evaluation procedure does not take this class into account. We evaluate the performances of the various algorithms using f_1-score for each category in the dataset.

B. Results

Table I shows a quantitative evaluation of the algorithms for semantic labeling. The structural context method gets the best overall classification rate and outperforms the baseline on 3 categories. This method is efficient to model the interactions of the superpixels of large areas like impervious surfaces or buildings. Objects with many superpixels are more likely to vote for the right label than objects with few supepixels because the error is divided between the neighbors. The poor performances of the model on cars is a consequence of the voting method. The superpixels of roads are more likely to consider a neighbor as a road than a car. The error is then propagated by the neighbors which lead the model to make a great number of bad predictions.

Figures 2 and 4 show the classification maps produced by the algorithms on tile 3 and 32. Figure 3 shows a zoom on a particular area of tile 3: the baseline model often confuses solar panels with cars while our structural model is able to correct this type of errors. Table II show how to read the classification maps: the colors in the table correspond to the colors in the classification maps. These results show that using structural context produces semantic maps with less noise caused by misclassifications of superpixels: it allows smart regularization of object borders.

IV. Conclusion

In the paper we described a context model for semantic labeling in aerial images. We use local graphs of interactions between superpixels of an image to model contextual relations. We use an efficient SSVM framework to learn a model of context with more than 100k training samples. We have shown it increases the classification performances on the challenging ISPRS dataset for urban semantic labeling.

TABLE I: f_1-score for each category in the dataset. The last column is the multiclass accuracy score.

Model	Imperv.	Build.	Veget.	Tree	Cars	Overall Acc.
			f_1-score			
Baseline 32	81.26	81.58	**62.71**	77.88	40.10	76.33
Baseline 64	81.13	82.36	62.46	76.13	**41.03**	75.98
Structural Context	**82.00**	**82.40**	58.18	**78.38**	32.46	**78.36**

V. Acknowledgment

The Vaihingen dataset was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html).

References

[1] D. J. Marceau, P. J. Howarth, J.-M. M. Dubois, and D. J. Gratton, "Evaluation Of The Grey-level Co-occurrence Matrix Method For Land-cover Classification Using Spot Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 513–519, 1990.

[2] S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 2, pp. 366–370, 2010.

[3] A. A. Popescu, I. Gavat, and M. Datcu, "Contextual Descriptors for Scene Classes in Very High Resolution SAR Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, pp. 80–84, jan 2012.

[4] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-Cover Mapping by Markov Modeling of Spatial-Contextual Information in Very-High-Resolution Remote Sensing Images," *Proceedings of the IEEE*, vol. 101, pp. 631–651, mar 2013.

[5] A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu, "Benchmarking classification of earth-observation data: from learning explicit features to convolutional networks," in *IEEE International Geoscience and Remote Sensing Symposium (Invited talk in the special session on Data Fusion)*, 2015.

[6] H. Randrianarivo, B. Le Saux, and M. Ferecatu, "Urban structure detection with deformable part-based models," in *International Geoscience and Remote Sensing Symposium*, 2013.

[7] A. Krizhevsky, I. Sulskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information and Processing Systems (NIPS)*, pp. 1–9, 2012.

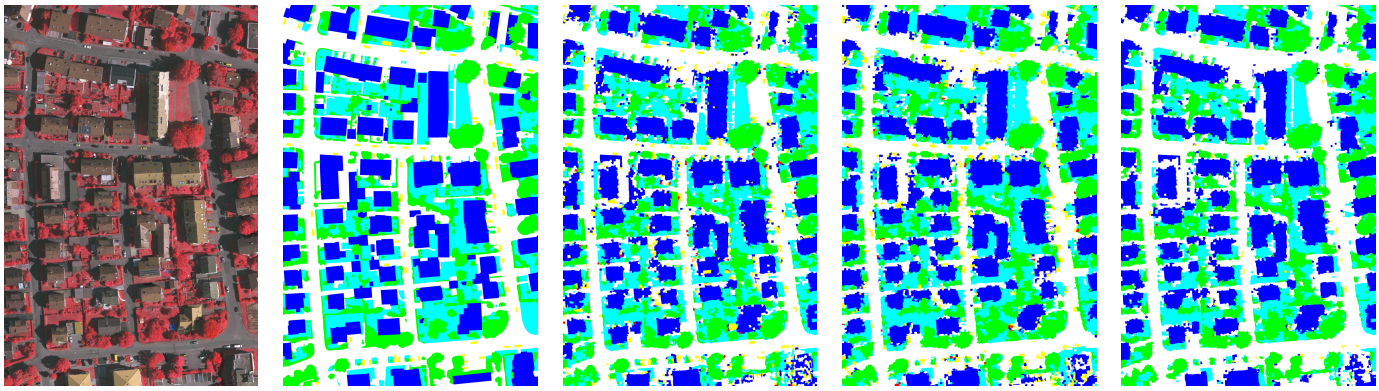
[8] R. Achanta, A. Shaji, and K. Smith, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *Pattern Analysis and . . .*, vol. 34, no. 11, pp. 2274–2281, 2012.

[9] S. Lacoste-julien, M. Jaggi, M. Schmidt, and P. Pletscher, "Block-Coordinate Frank-Wolfe Optimization for Structural SVMs," *International conference on Machine learning*, vol. 28, 2013.

[10] M. Gerke, J. Jung, C. Baillard, S. Benitez, G. Sohn, F. Rottensteiner, and U. Breitkopf, "the Isprs Benchmark on Urban Object Classification and 3D Building Reconstruction," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. I-3, no. September, pp. 293–298, 2012.

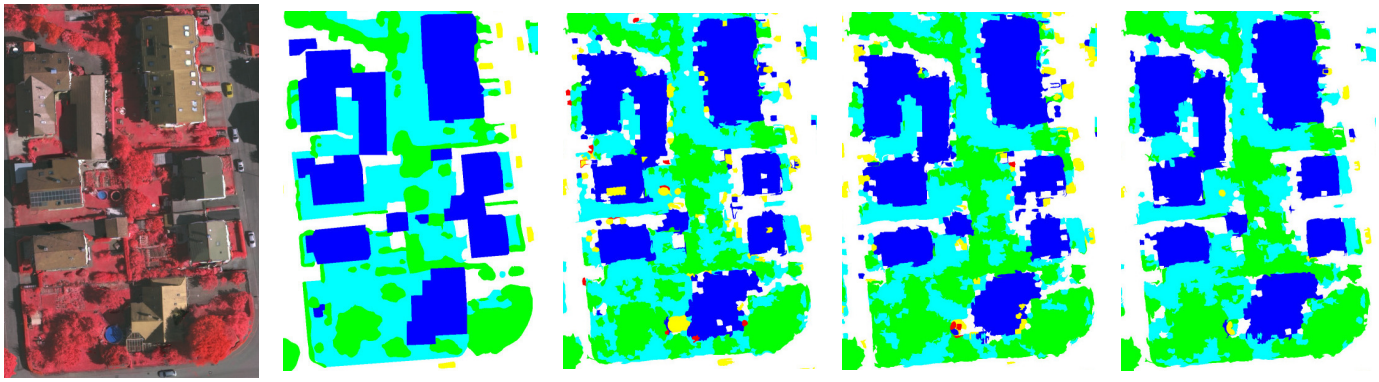
TABLE II: Ground truth classes for semantic labeling for the ISPRS dataset described in section III-A.

Impervious surface	Building	Low vegetation	Tree	Car	Clutter



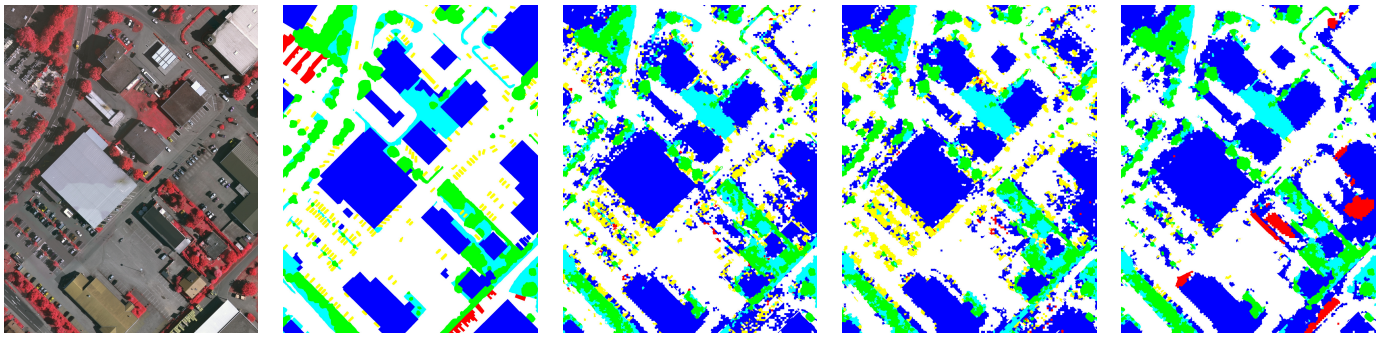
(a) Original image (b) Groundtruth (c) Baseline 32×32 (d) Baseline 64×64 (e) Structural context

Fig. 2: Semantic maps predicted for the tile 3



(a) Groundtruth (b) Groundtruth (c) Baseline 32×32 (d) Baseline 64×64 (e) Structural context

Fig. 3: Zoom on the tile 3. We observe that our method produces semantic maps with less noise than the baseline.



(a) Groundtruth (b) Groundtruth (c) Baseline 32×32 (d) Baseline 64×64 (e) Structural context

Fig. 4: Semantic maps predicted for the tile 32

doi: 10.2788/854791