GEOBIA MEETS PIXELS WITH HIERARCHICAL CONDITIONAL RANDOM FIELDS.

Devis Tuia and Michele Volpi

MultiModal Remote Sensing
University of Zurich, Switzerland

ABSTRACT

Land cover / land use classification of remotely sensed images is inherently geographical. The use of spatial information, accounting for neighborhood relationship and spatial smoothness of geographical objects, made its proofs in countless occasions and, especially when considering very high resolution images, methods ignoring spatial context do not perform well. Following this new requirement, two communities have appeared: one pursuing the traditional pixel-based image analysis and another considering image segmentation and classification (often referred to as Geographic Object-Based Image Analysis, or GEOBIA). Both communities have ignored each other, thus avoiding the striking synergies that could emerge from a joint use of the pixel-and object-based logics. In this paper, we propose an hybrid, hierarchical conditional random field model that enforces spatial smoothness and hierarchical consistency between the pixel and object-based maps. Using standard minimization tools, we show that letting the two spatial tessellations interact lead to strong improvements both in the numerical and visual sense.

Index Terms— Conditional random fields, very high resolution, random forests, Markov random fields, structured prediction, urban remote sensing.

1. INTRODUCTION

Very high resolution (VHR) imaging has driven recent research efforts in remote sensing: with the increased spatial resolution, VHR images permit to delineate objects precisely and to update land cover and land use maps timely. But to foster precise mapping, awareness of spatial context has become crucial [1], both in single and multimodal systems [2, 3].

But if the necessity of including information about spatial context meets general consensus, the way to perform such an inclusion still divides the community. Three main strategies are most commonly employed: 1) adding spatial filters in the input space [4], 2) segmenting the image into objects (i.e. the GEographic Object-Based Image Analysis, GEOBIA [5]) and 3) encoding spatial structure via structured models such as Markov [6] and conditional [7] random fields. These three types of strategies have been successful in their own way and therefore have mostly ignored each other ever since. Some exceptions exist, where authors mainly use segmentation to improve pixel-based products, for instance by using similarity measures accounting for both pixels and objects [8] or by extracting features from pixels and objects and using them in a common pixel-based classifier [9]: a real integration of the three strategies is still missing. Such an integration would be very beneficial, since all three strategies have complementary properties that could benefit from each other, if only a system was capable of handling pixel- and object-based logics, while leveraging their respective drawbacks. In this paper we aim at contributing towards this integration. In other words: to break the separation between the GEOBIA and pixel-based communities, which has no reason to exist.

We propose to use a hierarchical conditional random field to perform simultaneous pixel- and object-based image analysis. Our proposed hCRF model performs decision fusion between a pixel classification map and an object-based classifiction map by encoding 1) spatial structuring via contrast-sensitive priors [10] and 2) hierarchical consistency via hierarchical smoothing. The proposed system is casted in an efficient energy minimization framweork [11]. We apply the model on a multispectral very high resolution dataset composed of 20 urban scenes.

2. HIERARCHICAL CONDITIONAL RANDOM FIELD (HCRF)

With hCRF, we want to link two conditional random fields [12], one performing structured prediction at pixel level (p hereafter) on a pixel lattice $\mathcal{I}^p$ and another at object level (o hereafter) on a lattice $\mathcal{I}^o$ with as many entities as there are objects. To perform the joint optimization, we therefore propose a maximum and global a-posteriori decision rule to be maximized jointly over the pixel and object lattices (wrt $\mathcal{Y}$):

$$P(\mathcal{Y}|\mathcal{X}) \propto p(\mathcal{X}|\mathcal{Y})P(\mathcal{Y}^p,\mathcal{Y}^o) = p(\mathcal{X}|\mathcal{Y})P(\mathcal{Y}^p)P(\mathcal{Y}^o) \quad (1)$$

where $\mathcal{Y}$ are the random fields of class labels at the pixel ($\mathcal{Y}^p$) and object ($\mathcal{Y}^o$) levels, each one with C classes. $\mathcal{X}$ are the random fields of feature vectors for each level of granularity. We asusme conditional independence of feature vectors given labels and of father son relationships between objects

Thanks to XYZ agency for funding.

and pixels:

$$p(\mathcal{X}|\mathcal{Y}) = \prod_{s\in\{p,o\}} \prod_{i\in\mathcal{I}_s} p(x_i^s|y_i^s), \qquad P(\mathcal{Y}^p|\mathcal{Y}^o) = \prod_{i\in\mathcal{I}_p} P(y_i^p|y_\uparrow^o)$$

where $y_\uparrow^o$ is the label of the object to which the pixel belongs to at the object level and s stands for one of the two lattices (pixels or objects). We can now express the MAP decision rule (1) as an energy function to be minimized:

$$U(\mathcal{Y}|\mathcal{X}) = -\sum_{s\in\{p,o\}}\sum_{i\in\mathcal{I}_s} \ln \underbrace{p(x_i^s|y_i^s)}_{\text{[A]}} - \sum_{i\in\mathcal{I}_p} \underbrace{V(y_i^p, y_\uparrow^o)}_{\text{[B]}} + \sum_{s\in\{p,o\}} \underbrace{V'(Y^s|\mathcal{X}^s)}_{\text{[C]}} \quad (2)$$

where the three terms highlighted are:

[A] These are unary potential representing the posterior probability of a pixel (or object) $\mathbf{x}_i$ to belong to class y. Employing Bayes' theorem, it corresponds to a probabilistic output of a classifier. In the case of hCRF, two classifiers (one pixel-based and another object-based) are run in parallel. and the problemis casted as a problem with $2*C$ classes. By doing so, the unary potential are stacked as a block-diagonal matrix.

[B] A hierarchical pairwise term encouraging consistency among the two lattices (green line in Fig. 1). This term attributes minimal energy to the case, where the class predicted at the pixels level corresponds to the one predicted for the object the pixel belongs to at the object level. We use a Potts model between each child (pixel, y_i^p) and father (corresponding object $y_\uparrow^o$)

$$V(y_i^p, y_\uparrow^o) = 1 - \delta(y_i^p, y_\uparrow^o)$$

where δ is the Kronecker symbol.

[C] A pairwise term favouring spatial smoothness at each scale separately (red line in Fig. 1). This term is a CRF contrast sensitive potential [10] and penalizes different classes being predicted for neighbouring pixels (respectively objects) that look alike. In our case, we use a Gaussian kernel for contrast sensitivity: if the pixels are identical ($K(\mathbf{x}_i, \mathbf{x}_j) = 1$) and are predicted in different classes, penalization is maximal, while if they do not look alike ($K(\mathbf{x}_i, \mathbf{x}_j) = 0$) their prediction in different classes is not penalized

$$V'(y_i^s|\mathcal{X}^s) = \sum_{j\in\mathcal{N}_i} 1 - \delta(y_i^s, y_j^s) K(\mathbf{x}_i^s, \mathbf{x}_j^s)$$

where $\mathcal{N}_i$ is the spatial neighbourhood of sample i.

All the hierarchical (**B**) and pairwise (**C**) connection between pixels and objects are set in a single undirected graph that can then be used in any off-the-shelf energy minimization strategy. In this work, we used the Tree Reweighted Message Passing (TRWS) algorithm [11].

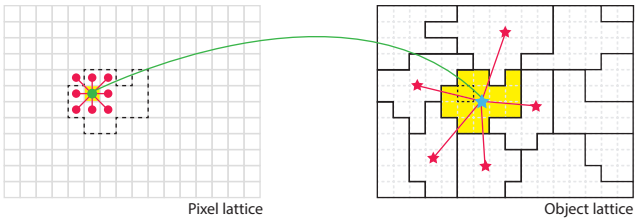


Fig. 1. General idea of the hCRF structured prediction model used to link the pixel and object-lattices. The green link represents the father-child potentials (**[B]** term in Eq. (2)). Red links represent the pairwise contrast-sensitive potentials (**[C]** term in Eq. (2)) between pixels (represented by circles) or objects centroids (represented by stars). The locations of the corresponding pixel (respectively superpixel) in the other lattice is highlighted with a dashed black line.

3. DATA, RESULTS AND DISCUSSION

3.1. Data and setup

We test the proposed hCRF on the Zurich summer dataset [13][1], composed of 20 pansharpened scenes acquired over the city of Zurich in 2002 (examples in Fig 2a and in the first column of Fig. 3). They are all part of the same QuickBird acquisition, but depict different types of urban areas and have different classes appearing. Manual annotation for eight classes is also available (see Fig. 3 for the color legend).

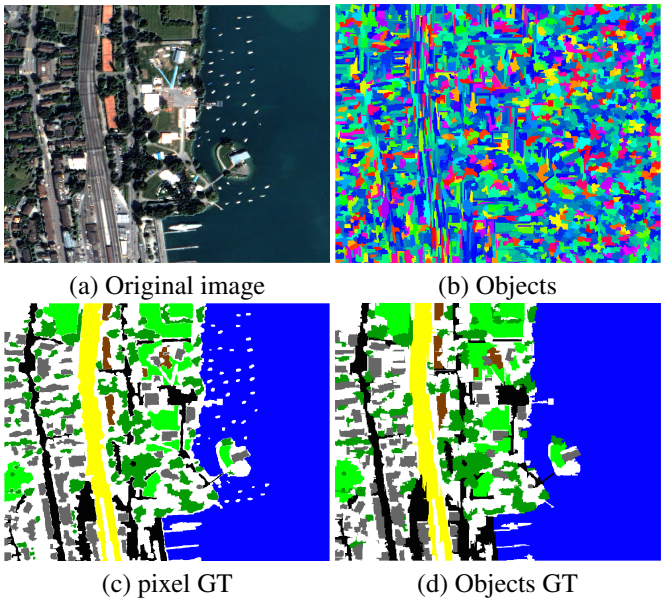
As stated above, we first perform two separate classification workflows at the pixel and object level.

- At the pixel level, we use color features (R-G-B-NIR raw values, normalized to standard scores), NDVI and NDWI indices and contextual features (average over 3×3 and 5×5 local windows, computed on the R and NIR bands). We have therefore a 10-dimensional input space.

- At the object level, we first extract superpixels using the Felzenwalb algorithm [14] and use them as objects. An example for image #3 can be seen in Fig. 2b. We used as features at the object level the min, max, average and standard deviation values of the pixels included in the object for the R, G, B, NIR, NDVI and NDWI channels. The object input space is therefore 24-dimensional. To train the models, we also derived an object ground truth by assigning the majority class found within the object (including background). An example of object ground truth can be seen in Fig. 2d

We train the two classifiers on images #1-15 and test on the remaining images #16-20. As classifiers, we used two random forests (RF). They are trained on 50% of the available pixels (thus xxx pixels) or objects (thus xxx objects).

[1]The dataset can be downloaded from `https://sites.google.com/site/michelevolpiresearch/data/zurich-dataset`



(a) Original image (b) Objects

(c) pixel GT (d) Objects GT

Fig. 2. Input data for image #3.

3.2. Results and discussion

Numerical results are reported in Tab. 1. At a first glance, the proposed hCRF boosts the results for each image and in both the pixel and object case. Even though all the classifiers considered are contextual (all the RF used use input space including contextual features either from local averages or statistics over the objects), the use of the hierarchical prior boosts the performances further, providing an increase in accuracy over the five test images of 3% (or 3 to 5κ points).

Figure 3 illustrates the maps obtained for each granularity level. We can appreciate the reduction in the salt and pepper labeling at the pixel scale, while at the object scale, we can observe switches to the correct label for the whole object (se, for instance, the railway class disappearing for image tile # 16). We also observe cases, where the hierarchical model switches the object label for an erroneous one (Figs. 3e-f): see, for instance, the river class for tile #17 or the bare soil patches in tile # 18.

Table 1. Numerical results on the five test tiles of the Zurich Summer dataset.

	Accuracy				Kappa			
Tile	Pixels		Objects		Pixels		Objects	
#	RF	hCRF	RF	hCRF	RF	hCRF	RF	hCRF
16	77.8	81.6	81.6	86.3	0.69	0.74	0.74	0.81
17	76.3	79.6	82.1	86.6	0.69	0.73	0.76	0.82
18	63.8	66.5	69.6	71.4	0.50	0.54	0.56	0.58
19	64.5	66.2	64.5	65.8	0.53	0.56	0.54	0.55
20	74.0	77.0	78.2	81.6	0.66	0.70	0.72	0.76
AVG*	72.7	**75.7**	76.7	**80.1**	0.66	**0.69**	0.70	**0.75**
Overall†	71.3	**74.2**	75.2	**78.4**	0.62	**0.65**	0.67	**0.71**

* average over the accuracies / kappas of the 5 test tiles.
† accuracies / kappas over all test samples.

4. REFERENCES

[1] M. Fauvel, Y. Tarabalka, J.A. Benediktsson, J. Chanussot, and J.C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, 2013.

[2] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[3] W. Liao, X. Huang, F. Van Collie, A. Gautama, W. Philips, H. Liu, T. Zhu, M. Shimoni, G. Moser, and D. Tuia, "Processing of thermal hyperspectral and digital color cameras: outcome of the 2014 data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 8, no. 6, pp. 2984–2996, 2015.

[4] D. Tuia, N. Courty, and R. Flamary, "Multiclass feature learning for hyperspectral image classification: sparse and hierarchical solutions," *ISPRS J. Int. Soc. Photo. Remote Sens.*, vol. 105, pp. 272–285, 2015.

[5] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photo. Remote Sens.*, vol. 65, no. 1, pp. 2–16, 2010.

[6] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial-contextual information," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 631–651, 2013.

[7] K. Schindler, "An overview and comparison of smooth labeling methods for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4534–4545, 2012.

[8] L. Z. Huo, P. Tang, Z. Zhang, and D. Tuia, "Semisupervised classification of remote sensing images with hierarchical spatial similarity," *IEEE Geosci. Remote Sens. Letters*, vol. 12, no. 1, pp. 150–154, 2015.

[9] W. Liao, F. Van Collie, F. Devriendt, S. Gautama, A. Pizurica, and W. Philips, "Fusion of pixel-based and object-based ffeature for classification of urban hyperspectral remote sensing data," in *Proc. GEOBIA*, Thessaloniki, Greece, 2014.

[10] P. Kohli, L. Ladicky, and P. H. Torr, "Robust higher order potentials for enforcing label consistency," in *Proc. CVPR*, 2008.

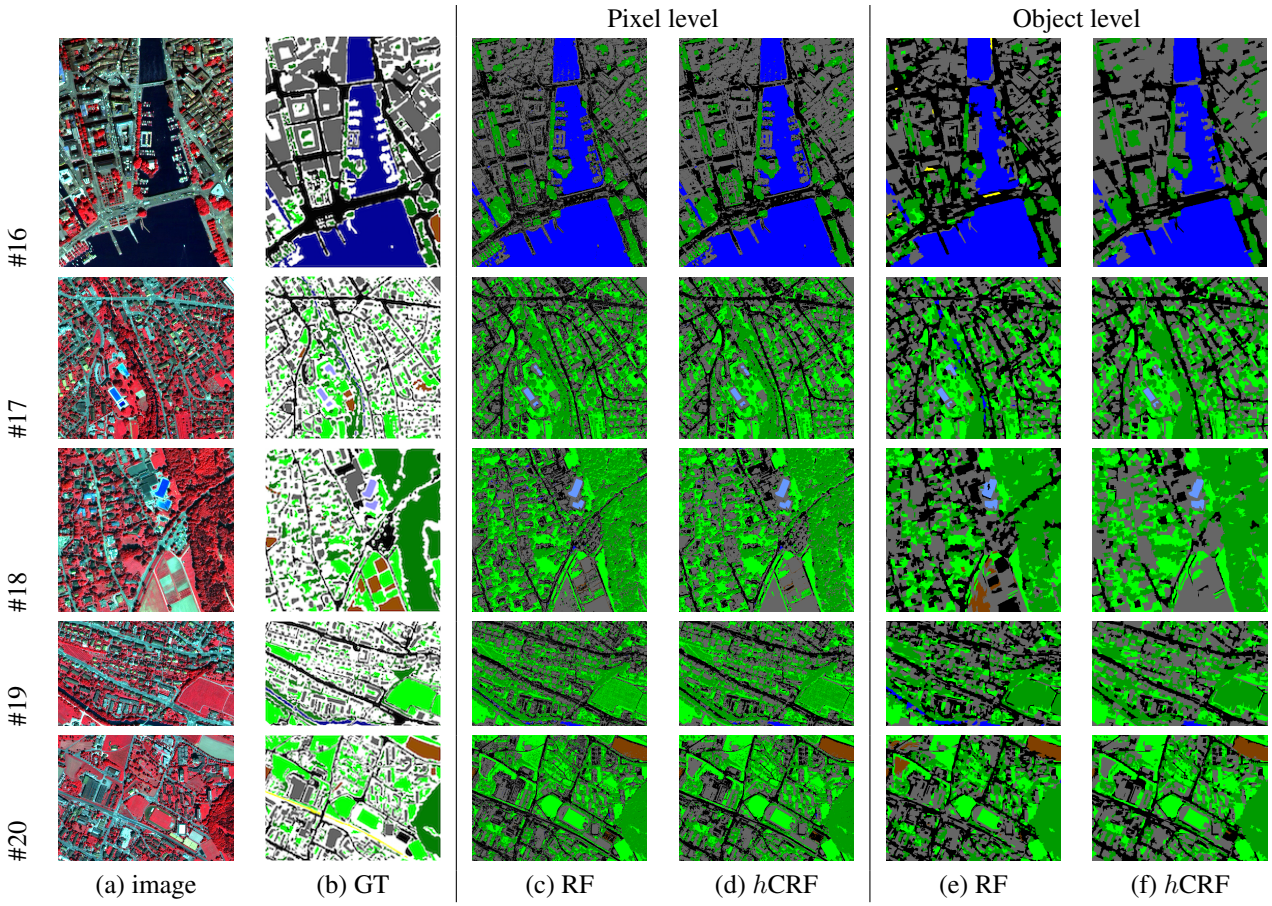


Fig. 3. Results on the five test images. (a) original image; (b) ground truth; (c) RF, pixel-based; (d) hCRF, pixel level; (e) RF, object-level; (f) hCRF, object-level. For the accuracies of the single maps, please refer to Tab. 1 (color legend: residential, street, trees, meadows, railway, water, swimming pools, bare soil).

[11] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, 2006.

[12] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov Random Fields," in *Proc. ECCV*, 2006.

[13] M. Volpi and V. Ferrari, "Structured prediction for urban scene semantic segmentation with geographic context," in *Proc. JURSE*, Lausanne, Switzerland., 2015.

[14] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, 2004.