IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9553253

SEMANTIC SEGMENTATION OF REMOTE SENSING IMAGES COMBINING HIERARCHICAL PROBABILISTIC GRAPHICAL MODELS AND DEEP CONVOLUTIONAL NEURAL NETWORKS

Martina Pastorino [1,2], *Gabriele Moser* [1], *Sebastiano B. Serpico* [1], *and Josiane Zerubia* [2]

[1] University of Genoa, DITEN dept., Genoa, Italy, martina.pastorino@edu.unige.it.
[2] Inria, Université Côte d'Azur, Sophia-Antipolis, France.

ABSTRACT

In this paper, a novel method to deal with the semantic segmentation of very high resolution remote sensing data is presented. Recent advances in deep learning (DL), especially convolutional neural networks (CNNs) and fully convolutional networks (FCNs), have shown outstanding performances in this task. However, the map accuracy depends on the quantity and quality of ground truth (GT) used to train them. At the same time, probabilistic graphical models (PGMs) have sparked even more interest in the past few years, because of the ever-growing need for structured predictions. The novel method proposed in this paper combines DL and PGMs to perform remote sensing image classification. FCNs can be exploited to deal with multiscale data through the integration with a hierarchical Markov model. The marginal posterior mode (MPM) criterion for inference is used in the proposed framework. Experimental validation is conducted on the ISPRS 2D Semantic Labeling Challenge Vaihingen dataset. The results are significant, as the proposed method has a higher recall than the standard FCNs considered and allows mitigating the impact of incomplete or suboptimal GT, especially with regard to the discrimination of minoritary classes.

Index Terms— CNN, FCN, PGM, Hierarchical Markov models, Semantic segmentation, Multiresolution images

1. INTRODUCTION

Models for multimodal data, typically based on multiview, multiscale, and multiresolution methods, are becoming increasingly important to face the requirements of remote sensing image processing [1]. Recent works have shown that DL techniques can reach very high per-pixel accuracies and even reproduce the correct shapes of the objects segmented. They are the dominating methods for image segmentation and have also gained increasing interest in remote sensing applications [2]. The most successful architectures are the FCNs [3], e.g., U-Net [4] and SegNet [5], which exhibit outstanding performances [6]. However, a major challenge is that DL architectures require big densely labeled GTs that accurately represent all object features, including – in particular – their boundaries. These fine-grained GTs are available only on benchmark datasets. At the same time, the interest in structured output learning and PGMs [7] has grown. Markov models postulated on planar or multilayer graphs are flexible and powerful stochastic models for spatial and multimodal information [8]. Two sub-classes of Markov models for 2-dimensional image analysis for which causality is formalized are Markov mesh random fields (MMRFs) on planar lattices [9] and hierarchical Markov random fields (MRFs) on quadtrees [10, 11]. For these two models, efficient inference algorithms can be employed. The two techniques are characterized by complementary properties: an MMRF captures spatial interactions among the pixels on a lattice and is a single-resolution model; a hierarchical MRF on a quadtree models dependencies among pixels located at different resolutions through a Markov chain, but does not explicitly characterize spatial dependencies within layers [10]. In a recent approach [12], the two strategies are combined and Markovianity is postulated both across the scales of a quadtree and with respect to the neighborhood system associated with each layer of the tree.

The information contained in multiresolution data guarantees accuracy and spatial precision of the classification maps [1] by exploiting the information at different resolutions: synoptic view and robustness to noise and outliers at the coarser resolutions and spatial detail at the finer ones. Indeed, the processing operations executed by a CNN [13] involve several multiscale processing stages, through convolutions and pooling operations, which intrinsically match the structure of multiresolution graph topologies on which PGMs can be efficiently formulated [7, 11].

In this paper, a novel method for semantic segmentation of multiresolution images based on hierarchical Markov models [8] and FCNs is proposed. Both the activations of the FCNs at different blocks (i.e., at several spatial resolutions) and the original image are used to build a training quadtree, and Markov chains are formulated both across the scales and with respect to a 1D scan of the pixel lattice of each layer. The model is combined with decision tree ensembles, such as Random Forest (RF) [14], to compute pixelwise posteri-

University of Genoa and Université Côte d'Azur (UCA) are part of the Ulysseus Alliance (European University). https://ulysseus.eu/

or probabilities necessary for the inference on the PGM with MPM, an especially advantageous criterion for classification tasks on multiresolution models [10]. The integration of these methodological components allows exploiting the representations extracted by the FCN across all its layers, incorporating prior information on the spatial behavior and the structure of the prediction output. This is aimed at mitigating the critical limitations of the FCN in learning spatial relations from sparse, incomplete, or suboptimal GTs, in which spatial class boundaries may not be present or may be poorly represented.

2. METHODOLOGY

2.1. Hierarchical Markov mesh random field

Let $\{S^0, S^1, \ldots, S^L\}$ be a set of pixel grids arranged as a quadtree: each site $s \in S^l$ has a parent site $s^- \in S^{l-1}$ and four children sites $s^+ \subset S^{l+1}$ $(l = 1, 2, \ldots, L-1)$. A hierarchy on the tree $S = \bigcup_{l=0}^{L} S^l$ from the root to the leaves is determined. If a discrete class label x_s in a finite set Ω of M classes ($x_s \in \Omega$, $s \in S$) is associated with each $s \in S$, then $\mathcal{X} = \{x_s\}_{s \in S}$ is a hierarchical MRF if [8, 10]:

$$P(\mathcal{X}^l | \mathcal{X}^{l-1}, \mathcal{X}^{l-2}, \ldots, \mathcal{X}^0) = P(\mathcal{X}^l | \mathcal{X}^{l-1}), \qquad (1)$$

where $\mathcal{X}^l = \{x_s\}_{s \in S^l}$ $(l = 1, 2, \ldots, L)$, i.e., if Markovianity holds across the scales. The model is extended to incorporate spatial information while maintaining the causality. Consider a rectangular lattice R and an order relation $\prec$ in the lattice of pixels, representing the pixels before each site $s \in R$ (i.e., the sites $r \in R$ such that $r \prec s$). A neighborhood relation is assumed in R consistently with this order relation, and $r \precsim s$ indicates that r is a causal neighbor of s. Hence, spatial Markovianity is expressed as:

$$P(x_s | x_r, r \prec s) = P(x_s | x_r, r \precsim s). \qquad (2)$$

A common choice is that the "past" of s is the set of all pixels traversed by a raster scan before s, whose neighbors are the three adjacent pixels located in the previous row and column. More details can be found in [9, 10, 11]. Here, (1)-(2) are assumed to hold jointly – with (2) being valid on each S^l ($l = 0, 1, \ldots, L$) –, defining a model where both the cross-layer and the intra-layer dependencies are characterized in order to deal with multiresolution images and capture spatial relations within each single-resolution layer, respectively.

2.2. DL architecture

DL has proved to be effective in the task of semantic segmentation, especially via FCNs [3], such as U-Net [4] or SegNet [5]. They use an encoder-decoder architecture in which, in addition to pooling layers where downsampling processes take place, the feature maps are also upsampled to match the original input resolution, therefore performing pixelwise predictions at the original resolution. In the present paper, this approach is extended to exploit the intrinsically multiscale behavior of CNNs through a hierarchical multiresolution MRF. The encoder is based on VGG16 [15].

The considered architectures have 5 convolutional blocks, each containing convolutional layers, zero paddings, followed by ReLU activations and batch normalizations. Each convolutional block is followed by a max pooling layer of size 2×2. The decoder, symmetrical to the encoder, performs the upsampling and the classification, learning how to restore the full spatial resolution while transforming the encoded feature maps into the final labels. The dimension of the patches used to train the network is 256×256 pixels. The loss function is computed by a pixelwise softmax [13] over the final feature map combined with the cross-entropy loss. Three skip connections, from three deconvolution blocks of the decoder to the output layer allow to collect the activations of the network at different resolutions, inserted into the quadtree to connect the FCN to the hierarchical PGM, in order to exploit the information hidden at different layers.

2.3. MPM inference and random forest

As pointed out in [11], the maximum *a-posteriori* estimate is not satisfying for multiscale image classification. The MPM criterion [7, 10] is especially appropriate for hierarchical MRFs because it penalizes errors according to the scale, avoiding error accumulation along the layers [10]. It can be proved that MPM on the proposed Markov model is accomplished through the following recursive steps [12];

$$P(x_s) = \sum_{x_{s^-}} P(x_s | x_{s^-}) P(x_{s^-}), \qquad (3)$$

$$P(x_s | y_s^d) \propto P(x_s | y_s) \prod_{t \in s^+} \sum_{x_t} \frac{P(x_t | y_t^d) P(x_t | x_s)}{P(x_t)}, \qquad (4)$$

$$P(x_s^c | x_s, y_s^d) \propto P(x_s | y_s^d) P(x_s | x_{s^-}) P(x_{s^-}) P(x_s)^{-n_s} \cdot \\ \cdot \prod_{r \precsim s} P(x_s | x_r) P(x_r), \qquad (5)$$

$$P(x_s | \mathcal{Y}) = \sum_{x_s^c} P(x_s^c | x_s, y_s^d) P(x_{s^-} | \mathcal{Y}) \prod_{r \precsim s} P(x_r | \mathcal{Y}), \qquad (6)$$

where y_s is the feature vector of site s, y_s^d collects the observations of all descendants of s in the tree (including s), x_s^c collects the labels of all sites connected to s (x_{s^-} and $\{x_r\}_{r \precsim s}$) and n_s is the number of such sites. First, (3) calculates $P(x_s)$ on all sites through a top-down pass from the root to the leaves. For the root layer, these probabilities are initialized as the relative frequency of the classes in the training set. Then, (4) and (5) compute $P(x_s | x_s^c, y_s^d)$ through a bottom-up pass from the leaves to the root. Finally, (6) derives $P(x_s | \mathcal{Y})$ through a second top-down pass, being $\mathcal{Y}$ the random vector of all feature vectors in the tree.

A symmetric visiting scheme on the pixel grid is used to prevent anisotropic artifacts, specifically a symmetrized com-

Table 1. Test-set results. Precision and recall are averaged over the classes. Per-class scores are recalls. OA and "Morpho" stand for overall accuracy and morphological erosion, respectively. Times include training and prediction.

	Architecture	**buildings**	**impervious**	**vegetation**	**trees**	**cars**	**OA**	**recall**	**precision**	κ	**F1**	**time** [s]
Full GT	Standard U-Net	**0.85**	**0.97**	0.42	0.84	0.80	**0.90**	0.78	**0.72**	**0.81**	**0.75**	5809
	Standard SegNet	0.83	0.96	0.40	0.81	0.79	0.89	0.76	0.70	0.78	0.72	5244
	Proposed, "PGM+NET" (U-Net)	0.80	0.90	**0.55**	**0.92**	**0.92**	0.86	**0.82**	0.60	0.74	0.69	11193
Morpho	Standard U-Net	0.91	**0.92**	0.15	0.62	0.27	**0.87**	0.57	**0.77**	0.75	0.66	5786
	Proposed, "PGM+NET" (U-Net)	0.89	0.73	0.49	**0.67**	0.62	0.76	0.68	0.55	0.60	0.61	11287
	Proposed, "Net for cars" (U-Net)	0.88	0.86	0.50	0.60	**0.70**	0.85	**0.71**	0.57	0.72	0.63	11271
	Proposed, "resize" (U-Net)	**0.93**	0.88	**0.51**	0.60	0.43	0.86	0.66	0.68	0.74	0.67	11324
	FESTA [16]	**0.93**	0.90	0.32	**0.67**	0.32	**0.87**	0.63	0.76	**0.77**	**0.69**	800332

bination of the zig-zag scan and the Hilbert curve, as explained in [12]. The transition probability $P(x_s|x_{s^-})$ across scales is defined by Bouman's model [17], i.e., $P\{x_s = \omega | x_{s^-} = \omega\} = \vartheta$ for all $\omega \in \Omega$, where ϑ is a parameter of the method, and $P\{x_s = \omega | x_{s^-} = \omega'\}$ is constant over all $\omega \neq \omega' (\omega, \omega' \in \Omega)$ [12]. The spatial transition probability $P(x_s|x_r)$ ($r \lesssim s$) is modeled analogously with a parameter ψ [12].

In the proposed method, the lattices S^l correspond to the various resolutions involved in the FCN, the observation vector y_s of each site $s \in S^l$ is obtained by stacking all network activations associated with that pixel location in the layers at resolution S^l, and RF [14] is used to estimate the pixelwise posteriors $P(x_s|y_s)$ from the training samples of the classes.

3. EXPERIMENTAL VALIDATION

The proposed method was experimentally validated with the ISPRS 2D Semantic Labeling Challenge Vaihingen dataset[1]. It consists of aerial images with resolution of 9 cm/pixel and six classes: buildings, impervious surfaces, low vegetation, trees, cars, and clutter. Indeed, clutter is highly mixed and of relatively limited interest as a target land cover class since it comprises all surfaces not included in the other classes. Each of 33 tiles includes near infrared, red, green bands and a digital surface model extracted from a LiDAR point cloud. Within the 16 images with "public" GT, 12 were chosen to train (tiles 1, 3, 7, 11, 13, 17, 23, 26, 28, 32, 34, and 37) and 4 to test the network (tiles 5, 15, 21, and 30). Experiments were run on Google Colab. Two subsections of 1024 × 1024 pixels of training tile 1 and test tile 5 were selected to train the RF and to apply the hierarchical PGM, respectively, since it was impossible to perform the analysis on bigger patches because of RAM limitations on Colab. These images lack instances of the class clutter, excluded from the experimentation.

The classification results shown in this paper were obtained with: $L = 4$, i.e., four levels in the quadtree, with a power-of-2 relation between layers; $\vartheta = \psi = 0.82$ (higher or lower values yielded worse results). Several training conditions were considered, since the Vaihingen dataset is an "ideal" one, with densely labeled GTs, normally unavailable in real-world scenarios. Therefore, the network was further trained with a "deteriorated" training set, in which part of the labeled pixels was removed randomly (shown in Fig. 1) or by morphological operators (see Table 1). This second approach was meant as an approximation of the GTs made of isolated patches of annotated labeled pixels usually found in real applications. In this case, however, the removal of labeled pixels was well balanced, preserving the prior probability. For brevity. the results are presented with U-Net as the baseline network, while SegNet is used as a benchmark applied to the full GT.

The pixelwise posterior probabilities are inserted in the hierarchical PGM, as mentioned before. However, the ones obtained by RF for class 5, "cars", on the finest lattice (1024 × 1024 pixels) would not be detailed enough because in this lattice there are no network activations and RF is applied only to the original image data (Fig. 1(d)). One way to overcome this problem is to directly use on the finest lattice the posteriors obtained in the output layer of the network ("PGM+NET"). Alternately, only focusing on the posteriors of class 5, the RF estimation can be replaced by either the network posteriors ("Net for cars") or a nearest neighbor upscaling of the posteriors from the coarser lattice 512 × 512 ("resize"). Amongst these three variants of the proposed approach, the rationale of the last two is to allow focusing on the discrimination of the minority classes (e.g., trees and cars in this dataset).

As it can be seen from the quantitative results reported in Table 1, the proposed approach exhibits remarkable improvements for the aforementioned classes, especially when the input training data approach the real-world GTs available for land-cover mapping applications. With the morphological erosion, for example, while the standard U-Net scored a recall of 0.27 in the classification of the cars, the proposed method was able to reach 0.70. Moreover, in all the considered situations, the recalls attained by the proposed approach were higher than those of the standard FCNs. In particular, in the above case "Net for cars", the gain in recall is about 14%.

The proposed technique was further compared to the recently proposed "FESTA" method, in which FCN training with a "scribbled" GT is addressed through an additional loss term that favors regularization in the spatial and feature domains [16]. The results suggest that FESTA also mitigated the impact of suboptimal GT. However, the proposed approach, across its variants, obtained higher or similar per-class recalls (higher especially for "cars") and slightly lower but similar

[1] https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/

OA and κ, with remarkably shorter computation times.

4. DISCUSSION AND CONCLUSION

A new approach for semantic segmentation of remote sensing images based on CNNs, hierarchical PGMs, and decision tree ensembles has been proposed in this paper. The reported results indicate that the proposed approach surpasses the accuracy of the standard FCN as per the recall. They suggest the capability of the proposed approach to exploit the spatial modeling ability of hierarchical Markov models to mitigate the limitations of FCN approaches in terms of training data requirements. The new method outperforms the state-of-the-art especially in the discrimination of minority classes, while maintaining adequate classification results for all classes.

Future work could involve the introduction of dense layers to compute the pixelwise posterior probabilities instead of the RF classifier. Furthermore, it would be interesting to combine the proposed method with transfer learning to favor prediction on a dataset different from the one it was trained on, in features and complexity, but with the same encoding of the classes, and compare its generalization performances to the ones of a standard FCN in the same framework.

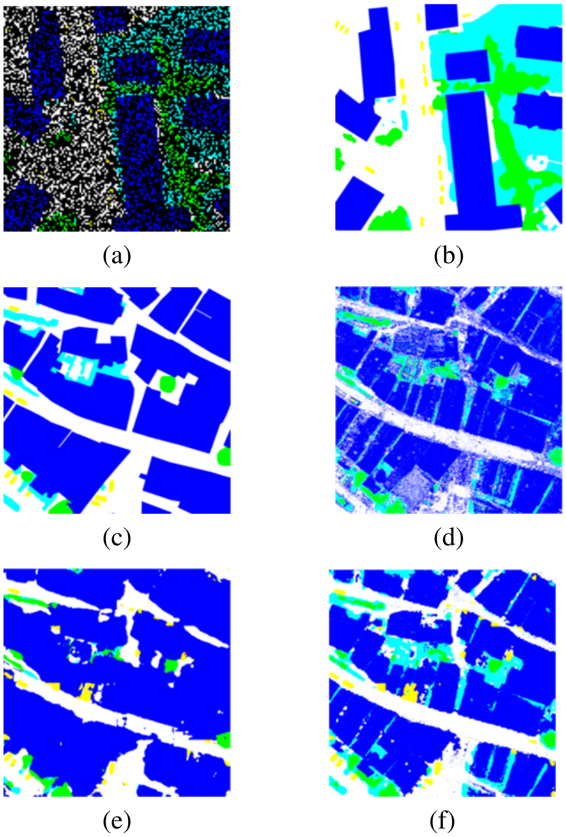

Fig. 1. **Ground truth and classification maps:** (a) training set with 70% of unlabeled pixels, (b) original training set, (c) test set; classification maps: (d) RF, (e) U-Net, and (f) the proposed method ("resize"). Classes: buildings (blue), impervious (white), vegetation (cyan), trees (green), cars (yellow).

5. REFERENCES

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

[2] K. Nogueira, O. Penatti, and J. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Patt. Rec.*, vol. 61, pp. 539–556, 2017.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3431–3440, 2015.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Med. Image Comput. Comput. Ass. Interv.*, ser. LNCS, vol. 9351. Springer, pp. 234–241, 2015.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[6] N. Audebert, B. Saux, and S. Lefèvre, "Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis.*, pp. 180–196, 2016.

[7] Z. Kato and J. Zerubia, "Markov random fields in image segmentation," *Found. Trends Signal Process.*, vol. 5, no. 1-2, pp. 1–155, 2012.

[8] S. Li, *Markov random field modeling in image analysis*, 3rd ed. Springer, 2009.

[9] P. A. Devijver, "Hidden Markov mesh random field models in image analysis," *J. Appl. Stat.*, vol. 20, pp. 187–227, 1993.

[10] J. Laferté, P. Pérez, and F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 390–404, 2000.

[11] I. Hedhli, G. Moser, S. B. Serpico, and J. Zerubia, "Classification of multisensor and multiresolution remote sensing images through hierarchical Markov random fields," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2448–2452, 2017.

[12] M. Pastorino, A. Montaldo, L. Fronda, I. Hedhli, G. Moser, S. B. Serpico, and J. Zerubia, "Multisensor and multiresolution remote sensing image classification through a causal hierarchical markov framework and decision tree ensembles," *Remote Sens.*, vol. 13, no. 5, 2021.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Boston, Massachussetts: USA: MIT Press, 2016.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv:1409.1556*, 2015.

[16] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geosci. Remote Sens. Lett.*, pp. 1–5, 2021.

[17] C. A. Bouman and M. Shapiro, "A multiscale random field model for bayesian image segmentation," *IEEE Trans. Image Process.*, vol. 3, no. 2, pp. 162–177, 1994.