IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883111

SEMANTIC SEGMENTATION OF SAR IMAGES THROUGH FULLY CONVOLUTIONAL NETWORKS AND HIERARCHICAL PROBABILISTIC GRAPHICAL MODELS

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

This paper addresses the semantic segmentation of synthetic aperture radar (SAR) images through the combination of fully convolutional networks (FCNs), hierarchical probabilistic graphical models (PGMs), and decision tree ensembles. The idea is to incorporate long-range spatial information together with the multiresolution information extracted by FCNs, through the multiresolution graph topology on which hierarchical PGMs can be efficiently formulated. The objective is to obtain accurate classification results with small datasets and reduce problems of spatial inconsistency. The experimental validation is conducted with several COSMO-SkyMed satellite images over Northern Italy. The results are significant, as the proposed method obtains more accurate classification results than the standard FCNs considered.

Index Terms— SAR imagery, semantic segmentation, CNN, FCN, PGM, hierarchical Markov models

1. INTRODUCTION

Current space missions allow satellite imagery to reach fine spatial resolutions. The data acquired can be optical (e.g., panchromatic, multispectral, and hyperspectral images) or radar, with different synthetic aperture radar (SAR) modalities (e.g., stripmap, spotlight, ScanSAR) and several trade-offs between resolution and coverage [1]. This offers great prospects for land cover mapping applications in the field of remote sensing. However, the development of a supervised method for the classification of SAR images – which suffer from speckle and are determined by complex scattering phenomena [2] – presents some issues.

Semantic segmentation methods based on deep learning (DL), for example fully convolutional networks (FCNs) [3, 4], are capable to reach accurate classification results, but may sometimes neglect spatial consistency during feature extraction [5], and they require large training datasets which are time consuming to retrieve [2]. The integration of information contained at different resolutions (e.g., the spatial details at finer resolutions and the robustness to noise and outliers at the coarser ones) guarantees more accuracy and spatial precision in the results of the segmentation [6]. In [7], for example, the semantic segmentation of SAR images is addressed with a multiscale convolutional neural network (CNN), which includes a feature concatenation stage to connect features with different scales and depths. Otherwise, in [8] an FCN has been modified with the addition of skip connections to integrate high-resolution feature maps from the early layers to obtain more accurate and detailed output classification maps. Multilevel feature fusion models such as [9] have also been considered to effectively integrate information belonging to features learned at different resolutions.

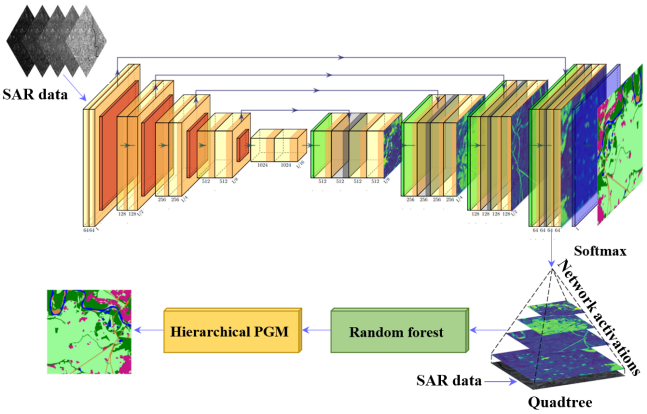

Fig. 1. Overall architecture of the method.

Indeed, the processing operations executed by CNNs (and, consequently, FCNs) [10] involve several multiscale processing stages, through which it is possible to obtain multiresolution information. Probabilistic graphical models (PGMs), such as Markov models, postulated on planar or multilayer graphs, are flexible and powerful stochastic models, capable to integrate spatial and multiresolution data [11, 12]. Therefore, their combination through a multilayer graph topology allows to incorporate information present at different resolutions.

In a recent approach [13], the two strategies (DL and stochastic models) are combined for the semantic segmentation of aerial images in case of scarce ground truth data. In the pre-

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

sent paper, this approach is extended to the semantic segmentation of SAR images through the combination of hierarchical Markov models [11] and FCNs. The integration of these methodological components aims to overcome the difficulties of the semantic segmentation of SAR images.

2. METHODOLOGY

The proposed approach (whose block diagram is shown in Fig. 1) aims to exploit the multiscale data representation extracted by a FCN to integrate with a hierarchical PGM through a quadtree topology. The idea is to train an FCN, characterized by an encoder-decoder architecture, with a dataset of SAR images and define a quadtree structure with a set of pixel grids S^l $(l = 0, \ldots, L)$ containing the network activations of the hidden layers of the decoder and the original bands of the input image. Activations at different convolutional layers correspond to different resolutions, thus allowing to incorporate multiscale information in the methodology.

Max pooling layers of dimension 2×2 are employed to match the power-of-2 relation which characterizes the pixel grids at different levels of a quadtree, in order to exploit the intrinsic multiscale nature of FCNs. Each site in the pixel grid $s \in S^l$ has a parent site $s^- \in S^{l-1}$ and four children sites $s^+ \subset S^{l+1}$ $(l = 1, 2, \ldots, L-1)$, and a hierarchy is defined on the tree $S = \bigcup_{l=0}^{L} S^l$ from the root to the leaves.

The hierarchical PGM defined on the quadtree is a combination between a hierarchical MRF on quadtrees, characterized by its causality and the capability to model multiresolution information, and a planar MRF, which can model spatial information between neighboring pixels. Therefore, the proposed method is able to capture both multiscale and spatial-contextual information. The proposed hierarchical PGM is defined by the following equations:

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

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

Equation (1) represents the Markovianity accross the different levels of the quadtree, $\mathcal{X}^l = \{x_s\}_{s \in S^l}$ $(l = 1, 2, \ldots, L)$, where x_s is the discrete class label of each pixel $s \in S$ in a set Ω of M classes. Spatial Markovianity is expressed in Equation (2) over each pixel grid S by a generic order relation $\prec$ representing the pixels $r \in S$ before each site $s \in S$ $(r \prec s)$. Since planar MRFs are generally non-causal, a neighborhood relation $r \precsim s$ is assumed in the pixel grid S to ensure causality at each scale in the quadtree. This relation is defined by a 1D scan of the pixel lattice based on zig-zag paths and Hilbert curves. This choice guarantees a symmetrical and causal pixel visiting, avoiding directional artifacts [14].

The marginal posterior mode (MPM) [12, 15] criterion is used for inference in the proposed multiscale approach. It is defined by three recursive steps in the quadtree [14] and it is especially advantageous for multiresolution graphs, as it is capable to penalize the errors according to the scale at which they occur, avoiding the accumulation of the errors along the layers [15]. This criterion maximizes the posterior probability of the label of each pixel, given all the observations in the quadtree.

Finally, to link the representation extracted by the FCN and the Bayesian inference structure of the PGM, an ensemble learning approach such as random forest (RF) [16] is integrated in the proposed method to compute a suitable set of pixelwise posteriors. The methodology is briefly summed up in Algorithm 1 and more details can be found in [13, 14].

Algorithm 1 FCN + MPM on the Hierarchical PGM

1: Training of the FCN with the input VHR dataset
2: Input to the MPM: $L-$levels quadtree containing, in the random field $\mathcal{Y} = \{y_s\}_{s \in S}$ of the observations, the network activations and the original channels of the image to classify
3: First top-down pass: estimation of the priors $P(x_s)$
4: Estimation of the posterior probabilities $P(x_s|y_s)$ through the RF classifier
5: Bottom-up pass: estimation of $P(x_s|y_s^d)$ and $P(x_s^c|x_s, y_s^d)$, where 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 $\{x_r\}_{r \precsim s}$)
6: Second top-down pass: estimation of $P(x_s|\mathcal{Y})$ at each level of the quadtree
7: Output: maximization of $P(x_s|\mathcal{Y})$

3. EXPERIMENTAL VALIDATION

The method was applied to a dataset of SAR images acquired by COSMO-SkyMed satellites in 2018 over Lombardy, Northern Italy. It consists of Stripmap GTC (Geocoded terrain corrected) images with a spatial resolution of 2.5 m and polarization HH. Since it is supervised, the proposed architecture requires ground truth data. The experiments were carried out referring to the DUSAF data archive (map "Land use and land cover of the Lombardy Region"), containing information for the year 2018, thus being consistent with the satellite imagery. Starting from the classification carried out in the DUSAF, five semantic macro-classes of interest were selected for the experiments: urban areas, low vegetation, tall vegetation (e.g., trees), bare soil, and water. Bare soil represents a negligible percentage of the pixels in the images and it is of relatively limited interest as a target land cover class, since it comprises several mixed surfaces (e.g., beaches, quarries, dumps, degraded areas, detrital accumulations).

The dataset consisted of tiles of size 2000×2000 multilooked at a resolution of 5 m to reduce the impact of the speckle, obtained by cropping and stacking 5 COSMO-SkyMed images collected during the winter (thus defining a short time series). Images acquired over the course of the other seasons were discarded to avoid possible land cover changes due to seasonal changes (e.g., volume of water bodies, seasonal ve-

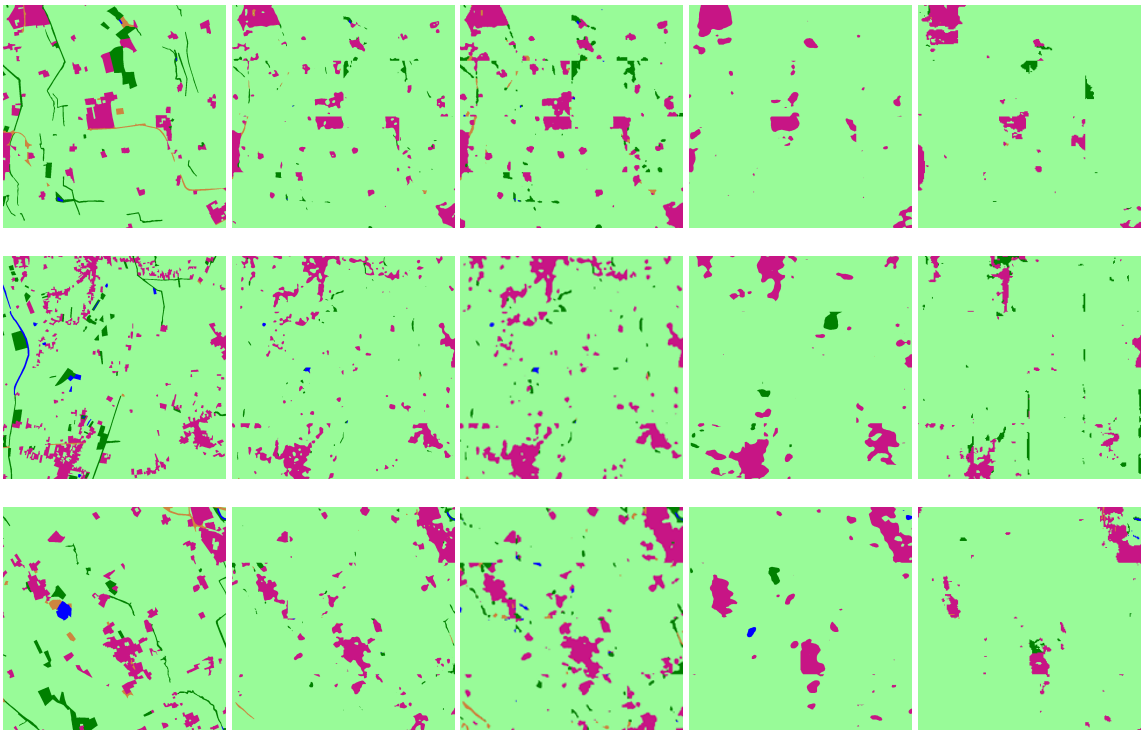

Fig. 2. **Ground truth and classification maps:** (a) ground truth and classification maps obtained with (b) FCN, (c) the proposed method, (d) LWN-Attention [17], and (e) HRNet [18]. Classes: urban (pink), low vegetation (pale green), trees (dark green), bare soil (orange), water (blue). (b) to (e): Product processed under a license of the Italian Space Agency (ASI); Original COSMO-SkyMed Product - ©ASI - (2018).

Table 1. Test-set results. Precision and recall are averaged over the classes. Per-class scores are recalls.

Architecture	urban	low vegetation	trees	bare soil	water	overall accuracy	recall	precision	F1 score
U-Net [19]	0.70	**0.99**	0.10	0.04	0.23	**0.92**	0.41	0.66	0.51
HRNet [18]	0.33	0.98	0.05	0.01	0.0	0.89	0.28	0.37	0.32
LWN-Attention [17]	0.45	**0.99**	0.01	0.01	0.01	0.90	0.29	0.33	0.31
Proposed ($\vartheta = 0.8, \psi = 0.4$)	**0.84**	0.95	**0.22**	**0.11**	**0.48**	0.91	**0.52**	0.46	0.49
Proposed ($\vartheta = 0.4, \psi = 0.8$)	0.76	0.98	0.11	0.04	0.25	**0.92**	0.43	**0.67**	**0.52**
Proposed ($\vartheta = 0.8, \psi = 0.8$)	**0.84**	0.95	**0.22**	0.10	0.38	0.91	0.50	0.46	0.48

getation). These images were used to train the FCN and RF and to test the proposed architecture combining FCNs and hierarchical PGMs. Experiments were run on an Alienware Aurora R11 with a RAM of 16 GB and a GPU NVIDIA GeForce RTX 2080 Ti.

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; and by optimizing through trial-and-error the values of the transition probability across the scales ϑ and of the spatial transition probability ψ. These parameters represent the probability that a site and its parent (or its causal neighbor) share the same label, therefore their value influence the modeling of spatial-contextual and multiresolution information in the hierarchical PGM. The results shown in Table 1 according to three combinations of ϑ and ψ (other results were omitted for brevity) demonstrate that incorporating more multiresolution information guarantees a better discrimination of the minority classes (e.g., water) but results in a small loss for the majority classes, while, fixing ϑ, lower values of the spatial transition probabilities yield a slight gain in recall.

The qualitative results shown in Fig. 2 suggest the effectiveness of the proposed method in discriminating the land cover classes considered, without an appreciable impact of the speckle on the classification output. In particular, even with a slight oversmoothing of the spatial edges between the classes, the method correctly identifies the urban areas in the scene. There is an underestimation of the samples belonging to the class "tall vegetation", which is interpreted as due to the overlap in the feature space with the class "low vegetation". As compared to the FCNs, the proposed method shows improvements both in the classwise and the average metrics.

The results obtained by the proposed technique were also compared with those of HRNet [18], a network consisting of multiresolution subnetworks connected in parallel, and of the light-weight attention network (LWN-Attention) in [17] (Fig. 2(d)-(e)). This technique is based on a multiscale feature fusion approach and makes use of multiscale information through the concatenation of feature maps at different scales [17]. The proposed method, leveraging both hierarchical and long-range information attained generally higher average classification results, thus suggesting that the proposed integration of FCN and hierarchical PGM can be advantageous.

4. DISCUSSION AND CONCLUSION

The semantic segmentation of SAR images through FCNs, hierarchical PGMs and decision trees ensembles has been addressed in this paper. The method aims to exploit the spatial and multiresolution modeling capabilities of hierarchical Markov models and FCNs to obtain accurate classification results. The experimental validation on COSMO-SkyMed satellite images demonstrated the potential of this approach, which generalizes a previous technique developed in the framework of aerial optical imagery, for SAR data. The reported results indicate that the employed method surpasses the classification accuracy of the FCNs, in particular in the discrimination of minority classes. Furthermore, the classification maps generated by the algorithm are quite visually regular, but without exhibiting spatial oversmoothing.

Future work may involve the extension of the proposed model to the multisensor case, for example with optical or SAR data acquired by distinct missions with different radar bands and spatial resolutions.

5. ACKNOWLEDGMENT

Project carried out using COSMO-SkyMed Products, © of the Italian Space Agency (ASI) delivered under a license to use by ASI. The activity of the first three authors was partially supported by ASI in the framework of the project MultiBigSARData - ASI no. 2021-7-U.0; the support is gratefully acknowledged.

6. REFERENCES

[1] J. A. Richards, *Remote sensing digital image analysis: an introduction*, 5th ed. Springer, 2013.

[2] W. Wu, H. Li, X. Li, H. Guo, and L. Zhang, "PolSAR image semantic segmentation based on deep transfer learning—realizing smooth classification with small training sets," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 977–981, 2019.

[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] S. Šćepanović, O. Antropov, P. Laurila, Y. Rauste, V. Ignatenko, and J. Praks, "Wide-area land cover mapping with sentinel-1 imagery using deep learning semantic segmentation models," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 10 357–10 374, 2021.

[5] F. Ma, F. Gao, J. Sun, H. Zhou, and A. Hussain, "Weakly supervised segmentation of SAR imagery using superpixel and hierarchically adversarial CRF," *Remote Sens.*, vol. 11, no. 5, 2019.

[6] 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.

[7] Y. Duan, X. Tao, C. Han, X. Qin, and J. Lu, "Multi-scale convolutional neural network for SAR image semantic segmentation," in *2018 IEEE GLOBECOM*, pp. 1–6, 2018.

[8] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 223–236, 2019.

[9] R. Zhang, J. Chen, L. Feng, S. Li, W. Yang, and D. Guo, "A refined pyramid scene parsing network for polarimetric SAR image semantic segmentation in agricultural areas," *IEEE Geosci. and Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

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

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

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

[13] M. Pastorino, G. Moser, S. B. Serpico, and J. Zerubia, "Semantic segmentation of remote sensing images combining hierarchical probabilistic graphical models and deep convolutional neural networks," in *2021 IGARSS*, pp. 8672–8675, 2021.

[14] 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.

[15] 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.

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

[17] S. Liu, C. He, H. Bai, Y. Zhang, and J. Cheng, "Light-weight attention semantic segmentation network for high-resolution remote sensing images," in *2020 IGARSS*, pp. 2595–2598, 2020.

[18] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 IEEE CVPR*, pp. 5686–5696, 2019.

[19] 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.