Learning Deep Features on Multiple Scales for Coffee Crop Recognition

Rafael Baeta[1], Keiller Nogueira[1], David Menotti[2], Jefersson A. dos Santos[1]

[1]*Department of Computer Science, Universidade Federal de Minas Gerais, Brazil*
{rbaeta,keiller.nogueira,jefersson}@dcc.ufmg.br,
[2]*Department of Informatics Federal University of Paraná, Brazil*
Email: menotti@inf.ufpr.br

***Abstract*—Geographic mapping of coffee crops by using remote sensing images and supervised classification has been a challenging research subject. Besides the intrinsic problems caused by the nature of multi-spectral information, coffee crops are non-seasonal and usually planted in mountains, which requires encoding and learning a huge diversity of patterns during the classifier training. In this paper, we propose a new approach for automatic mapping coffee crops by combining two recent trends on pattern recognition for remote sensing applications: deep learning and fusion/selection of features from multiple scales. The proposed approach is a pixel-wise strategy that consists in the training and combination of convolutional neural networks designed to receive as input different context windows around labeled pixels. Final maps are created by combining the output of those networks for a non-labeled set of pixels. Experimental results show that multiple scales produces better coffee crop maps than using single scales. Experiments also show the proposed approach is effective in comparison with baselines.**

***Index Terms*—Deep Learning; Remote Sensing; Coffee Crops; High-resolution Images; Agriculture.**

I. Introduction

Cropland planning, which is typically represented by using thematic maps, is fundamental in computational agribusiness applications. In some countries, the correct mapping of crop areas is also a requirement for including producers in government funding programs. Although many land use surveys are still done manually, the use of remote sensing images as a source of information to automatically create thematic maps is becoming more common over the years [1, 2].

Despite of recent advances in image acquisition and in supervised classification algorithms, recognition of crop regions in remote sensing images still poses many challenges. In the State of Minas Gerais (Brazil), coffee farming is a very important economic activity [3]. Coffee crop recognition is difficult because it is usually cultivated in mountainous regions. This causes shadows and distortions in the spectral information, which makes hard the classification and interpretation of shaded objects in the image because spectral information is either reduced or totally lost. Moreover, growing of coffee is not a seasonal activity, and, therefore, in the same region, there may be coffee plantations of different ages, which also affects the observed spectral patterns [4, 5].

Traditional automatic methods based on supervised classification for high spatial resolution remote sensing images are composed by three main steps [1, 2, 6]: (i) segmentation, (ii) feature extraction and, (iii) training. In this context, many works have been demonstrated the importance of combining features from multiple scales in order to obtain high quality automatic thematic maps [6, 7].

More recently, new state-of-the-art results have been achieved by deep learning-based approaches in pattern recognition for remote sensing applications [8]. The main advantages of deep-based strategies is the capability of learning data-driven spatial features and classifiers (in different layers) and adjusting this learning, in running time, based on the accuracy of the network, giving more importance to one layer than another depending on the problem. As a drawback, these deep learning strategies usually require too much labeled samples. Concerning automatic creation of thematic maps, specifically, some strategies have been proposed for learning not only high-quality spectral-spatial features but also the pixel context in an integrated way, which avoid the need of segmenting the image as a first step [9, 10].

This work aims at identifying coffee plantations in high spatial resolution remote sensing images. The proposed approach employs deep-based semantic segmentation to automatically learn coffee patterns. In order to obtain more accurate maps, we have proposed to learn and combine coffee patterns in multiple scales, as presented in Fig. 1. More specifically, we proposed and evaluated three convolutional neural network architectures based on *context window concept* [9] and a majority voting scheme to combine their outputs in a single multiscale final map.

The remainder of this paper is structured as follows. Related work is presented in Section II. Section III presents the methodology. Experimental protocols as well as obtained results are discussed in Section IV. Finally, in Section V we conclude the paper and point out promising directions for future work.

II. Related Work

The development of algorithms for spatial extraction information is a hot research topic in the remote sensing

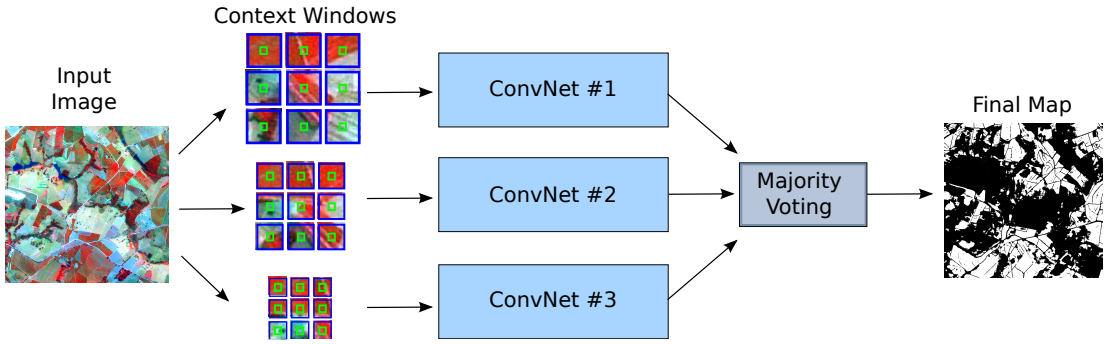

Fig. 1. **The proposed approach.** Non-labeled pixels are classified by ConvNets that consider different-size *context windows* around each pixel as input. This process generates a probability map for each ConvNet that are combined to create the Final Map.

community [11]. It is mainly motivated by the recent accessibility of high spatial resolution data provided by new sensor technologies. Even though many visual descriptors have been proposed or successfully used for remote sensing image processing [12–14], some applications demand more specific description techniques. As an example, very successful low-level descriptors in computer vision applications do not yield suitable results for coffee crop classification, as shown in [15]. Thus, common image descriptors can achieve suitable results in most of applications. Furthermore, higher accuracy rates are yielded by the combination of complementary descriptors that exploits late fusion learning techniques. Following this trend, many approaches have been proposed for selection of spatial descriptors in order to find suitable algorithms for each application [16–18]. Cheriyadat [17] proposed a feature learning strategy based on Sparse Coding, which learned features from well-known datasets are used for building detection in larger image sets. Faria et al. [16] proposed a new method for selecting descriptors and pattern classifiers based on rank aggregation approaches. Tokarczyk et al. [18] proposed a boosting-based approach for the selection of low-level features for very-high resolution semantic classification.

Despite the fact the use of Neural Network-based approaches for remote sensing image classification is not recent [19], its massive use is recent motivated by the study on deep learning-based approaches that aims at the development of powerful application-oriented descriptors. Many works have been proposed to learn spatial feature descriptors [20–23]. Firat et al. [20] proposed a method that combines Markov Random Fields with ConvNets for object detection and classification in high-resolution remote sensing images. Hung et al. [21] applied ConvNets to learn features and detect invasive weed. In [22], the authors presented an approach to learn features from Synthetic Aperture Radar (SAR) images. Zhang et al. [23] proposed a deep feature learning strategy that exploits a pre-processing salience filtering. Moreover, new effective hyperspectral and spatio-spectral feature descriptors [24–27] have been developed mainly boosted by the deep learning growth in recently years.

Regarding coffee crop recognition tasks, some recently published works have proposed approaches based on supervised classification to classify segmented regions [4, 6]. Other works have also addressed the coffee crop recognition problem by improving scene classification instead of investigating semantic segmentation [28, 29]. Finally, some effort has been done in order to apply deep-based strategies to create coffee land-cover maps [5, 9]. Besides the aforementioned efforts, to the best of our knowledge, the proposed approach differs from literature because there is no other one for geographical mapping of coffee crops based on deep learning considering multiple pixel-wise semantic segmentation scales for high-resolution remote sensing images.

III. The Proposed Approach

Our approach consists in the combination of Convolutional Networks (ConvNets) that learn patterns in different scales to assign a class to each pixel from an input remote sensing image. This strategy is based on the notion of *context windows* proposed in [9] and introduced in Section III-A. These windows are used by the ConvNets, which architecture is presented in Section III-B, to extract information and classify the pixels. Finally, the final multi-classifier is described in Section III-C.

A. Context Windows

As introduced, the proposed coffee crop mapping is based on a pixel-wise technique, in which each and every pixel of the input image is classified independently. Given that the information extracted from the pixel itself may not be enough to allow its classification, we employ the notion of *context window*. This notion is based on the fact that the pattern of each pixel is represented by a sufficiently large *context window* which is centered on the pixel in order to include the pattern of its neighborhood. This window allows the approach to extract relevant information about the region of the centered pixel, which may help in its final classification. Note that the *context window* may require a different ConvNet architecture depending on its size to capture the relative scale patterns. Thus, it is obviously that different size of context windows (one for each scale) produces distinct features for the same pixel.

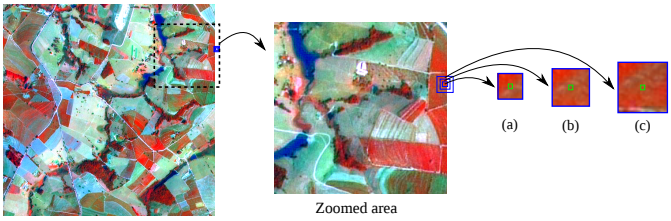

Fig. 2. Example of multiple context windows for the same pixel (a–c). The patterns are represented by windows centered on the pixel of interest to include the context of its neighborhood.

B. Convolutional Networks Architecture

Convolutional networks (ConvNets) [30], a deep learning technique, are defined by neural network architectures typically composed of many layers. Each layer is composed of processing units, also known as neurons. They can simultaneously learn data-driven features and classifiers. Furthermore, the learning rate can be adjusted, in runtime, based on the accuracy of the network.

This feature learning step may be stated as a technique that learns a transformation of raw data input to a representation that improves the class separability [30]. Since encoding the spatial features in an efficient and robust fashion is the key for generating discriminatory models, the feature learning step is a great advantage of ConvNets when compared to conventional methods, such as low- and mid-level methods. This advantageous process takes place in multiple layers (responsible for encoding spatial features automatically) that learn adaptable and specific feature representations in a data-dependent hierarchical way. As a result, low-level descriptors are learned in initial layers of the network and high-level features in the deeper ones. This process aims to extract all feasible information from the data, which creates robust features and classifiers.

Formally, given a set of labeled sample pixels and their contextual windows, the process for learning representations and classifiers to semantically segment remote sensing images consists in training a ConvNet to learn the feature patterns that compose the class of interest regions.

This process was performed, in this paper, by three networks presented in Fig. 3, Fig. 4, and Fig.5. It is important to point out that the ConvNet architecture is dependent on the contextual window size. Complex patterns with many objects and structures may require large window size. Consequently, large window size requires more complex ConvNets, i.e., more layers, filtering and pooling operations. The architectures, based on the one proposed by Nogueira et al. [9], have three convolutional layers and two fully-connected. However, they differ considerably on the size of filters and the strides of the convolutional layers.

Specifically, the *ConvNet #1* receives as input 17×17 pixels context windows and all convolutional layers are composed by 3×3 filters with stride 1. The *ConvNet #2* receives as input 25×25 pixels context windows. The first two convolutional layers are composed by 4×4 filters with stride 1. The last convolutional layer optimize 3×3 filters with stride 1. The *ConvNet #3* receives as input 33×33 pixels context windows and all convolutional layers are composed by 4×4 filters with stride 1.

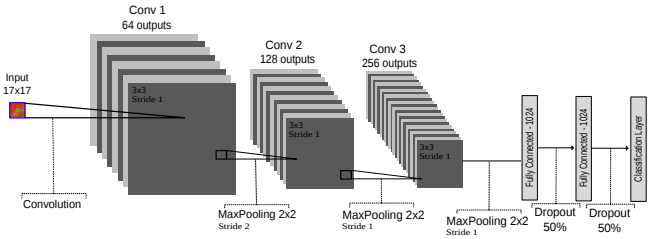

Fig. 3. ConvNet #1: architecture with 17×17 context windows as input.

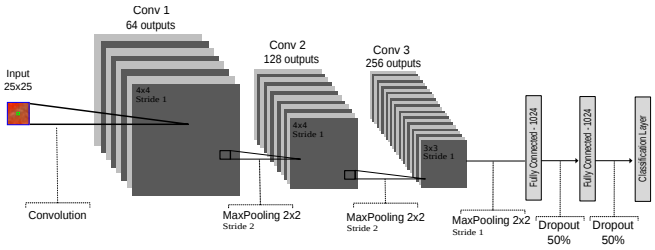

Fig. 4. ConvNet #2: architecture with 25×25 context windows as input.

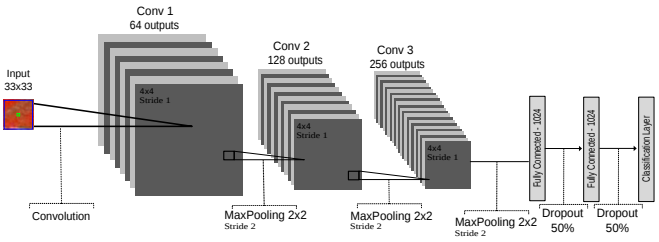

Fig. 5. ConvNet #3: architecture with 33×33 context windows as input.

As can be seen in Fig. 3, Fig. 4, and Fig. 5, some auxiliary techniques were employed between some of layers, such as dropout regularization [31] and max pooling. It is important to emphasize that Rectified Linear Unit (ReLU) is the processing unit selected to be used in all layers of the proposed ConvNets because of its advantages when compared to others (such as Hyperbolic Tangent and Sigmoid), including: (i) works better to avoid saturation during the learning process; (ii) induces the sparsity in the hidden units; and (iii) does not face gradient vanishing problem [32] as with Sigmoid and Hyperbolic Tangent functions.

The size of context windows are empirically defined according with the usual size of coffee crops in real world. Thus, *ConvNet #1* considers $20m$ around each pixel, *ConvNet #2* encodes features in a $30m$ radio, and *ConvNet #3* considers $40m$ around the pixels. The main difference among them is the classical tradeoff between context information and noise, i.e., the largest ConvNet is used to encode more context and

less noise, while the small one may be used in cases when larger context brings too much noise, disturbing the results.

C. Multi-scale Classifier

Given an input image, the process of creating a coffee crop map consists in two main steps: (i) classification of context windows; and (ii) the combination of probabilities maps from each ConvNet.

As mentioned, the first step is the classification of the multiple context windows for each unlabeled pixel by using the ConvNets proposed in Section III-B. This process also works for a set of non-contiguous pixels, which means that predicted regions could be in the same remote sensing image. When a context window is classified by a ConvNet, the probability function generated by this classification is, in fact, associated with the pixel at the center of that window. This process allows the method to create a probability map over entire regions (or images), which, after some post-processing method, results in a semantic segmented image.

The second step is the combination of the output probability map from each ConvNet. In this work, we combine the output probability maps from the ConvNets by using a majority voting scheme, as shown in Fig. 6. Given the class probabilities of each ConvNet, the final class is defined by their sum. The final class is the one with maximum sum probability.

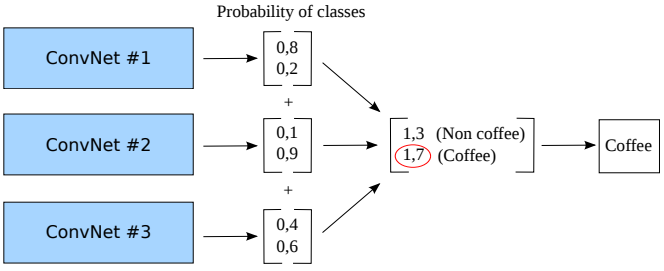

Fig. 6. Example of probability combination for deciding the final class of an input pixel.

IV. Experimental Evaluation

In this section, we present the experiments that we performed to validate our method. We have carried out experiments in order to address the following research questions: (1) is multiple scale combination more effective than individual ConvNets for semantic segmentation of coffee crops? (2) Are the proposed methods effective in the coffee crop recognition problem when compared to the baselines?

A. Setup

1) Baselines: We compare the proposed method against two approaches that follows the traditional three-main-step strategy: (i) segmentation, (ii) feature extraction and, (iii) classification. These approaches, named here as *MSC-Boost* and *HMSC-Boost*, are based on boosting of classifiers and combine features from multiple segmentation scales [6]. In our experiments, both approaches are implemented to consider features extracted from five segmentation scales. The main difference between them is that *MSC-Boost* consider all regions segmented over the segmented scales while *HMSC-Boost* starts from the coarse regions and use the other scales in sequence as refinement steps. We have used the same engineered features of the original paper [6]. For a better comparison, we also include results with a SVM with RBF and the best engineered descriptor in the best segmentation scale as reported in [6].

2) Dataset: It is a composition of scenes taken by the SPOT sensor in 2005 over Monte Santo de Minas county, State of Minas Gerais, Brazil. This area is a traditional place of coffee cultivation, characterized by its mountainous terrain. In addition to common issues in the area of pattern recognition in remote sensing images, these factors add further problems that must be taken into account. In mountainous areas, spectral patterns tend to be affected by the topographical differences and by interferences generated by shadows. This dataset provides an ideal environment for multi-scale analysis, since the variations in topography require the cultivation of coffee in different crop sizes. Another problem is that coffee is not an annual crop. This means that, in the same area, there are plantations of different ages. In terms of classification, we have several completely different patterns representing the same class while some of these patterns are much closer to other classes. The dimensions of the image used are 3000×3000 pixels with spatial resolution equals to $2.5m$. To facilitate the experimental protocol, we divided the dataset into a grid of 3×3, generating 9 subimages with dimensions equal to 1000×1000 pixels. In the experiments, we used 10 different sets of 1 million pixels each, to be used for training and classification (testing stage). The results of the experiments described in the following sections are obtained from all combinations of the 9 subimages used (6 for training and 3 for classification).

3) Assessment of results: To analyze the results, we computed the overall accuracy and Kappa index for the classified images at each iteration. In our experiments, the overall accuracy is defined as the sum of true positive and true negative samples divided by the total number of samples. Kappa is an effective index to compare classified images, commonly used in RSI classification [6]. Experiments in different areas show that Kappa could have various interpretations and these guidelines could be different depending on the application. However, Landis and Koch [33] characterize Kappa values over 0.80 as "almost perfect agreement", 0.60 to 0.79 as "substantial agreement", 0.40 to 0.59 as "moderate agreement", and below 0.40 as "poor agreement". Negative Kappa means that there is no agreement between classified data and verification data.

4) Implementation details: The proposed approach was implemented by using the Tensorflow framework. This framework is more suitable due to its support to parallel programming using CUDA, a NVIDIA parallel programming based on graphics processing units. Therefore, Tensorflow adopted and employed along with libraries as CUDA and CuDNN4. The complete set of experiments was performed on a 64 bits

Intel i7 4960X machine with 3.6GHz of clock and 64GB of RAM memory. We have used the following GPUs: a GeForce GTX770 with 4GB of internal memory and a GeForce GTX Titan X with 12GB of memory, both under a 7.5 CUDA version. Ubuntu version 14.04.3 LTS was used as operating system. The ConvNet and its parameters were adjusted by considering a full set of experiments based on [9].

B. Results

1) Multiple × Individual Scales: In this section, we compare the classification results obtained by using individual scales represented by *ConvNet #1*, *ConvNet #2*, and *ConvNet #3* against the combination of scales by using the proposed combination scheme. Table I presents the classification results.

TABLE I
CLASSIFICATION USING CONVNETS OVER DIFFERENT SCALES AND THE COMBINED RESULTS.

Scale	Overall Acc. (%)	Kappa (κ)
#1	87.57 ± 1.58	0.713 ± 0.016
#2	88.40 ± 1.27	0.725 ± 0.017
#3	88.02 ± 1.47	0.719 ± 0.023
Combination	88.90 ± 4.00	0.739 ± 0.054

According to the results, one can observe that the combination of scales achieves better maps than the best individual scale. We can suppose that the proposed combination improves the results by exploiting the diversity of individual ConvNets in different scales.

Overall, we could observe that the voting scheme combination create an intermediate result among each scale, as expected. We show an example of result for each single scale and the combination of them in Fig. 7. Observe, for instance, the reduction of false positives pixels (red) from each scale in comparison with the combination map.

2) Comparison to the baselines: In Table II is presented the results for the proposed approach and the baselines.

TABLE II
CLASSIFICATION RESULTS COMPARING THE PROPOSED APPROACH AGAINST THE BASELINES.

Approach	Overall Acc. (%)	Kappa (κ)
SVM (RBF)	80.09 ± 1.58	0.748 ± 0.025
MSC-Boost	82.28 ± 1.60	0.780 ± 0.025
HMSC-Boost	82.69 ± 1.68	0.788 ± 0.024
Ours	88.90 ± 4.00	0.739 ± 0.054

Concerning overall accuracy, one can be note the proposed approach overcome the results of the baselines. This shows that the combination of ConvNets on different scales can be a powerful tool for recognition of coffee crops. On the other hand, obtained results are not competitive observing Kappa index.

There are two important issues that may justify this results and can lead us to future improvements: (1) ConvNets are very sensitive to unbalanced dataset and deep features learned may not be enough to represent coffee crops; and (2) fully-trained ConvNets does not perform well for small training datasets. In this sense, obtained results match the conclusion presented by [29], in which BIC descriptor outperforms fully-trained ConvNets in different configurations. The best way to improve results is to perform fine tuning over a pre-trained network. The use of more powerful classifiers instead of softmax also leads to improvements.

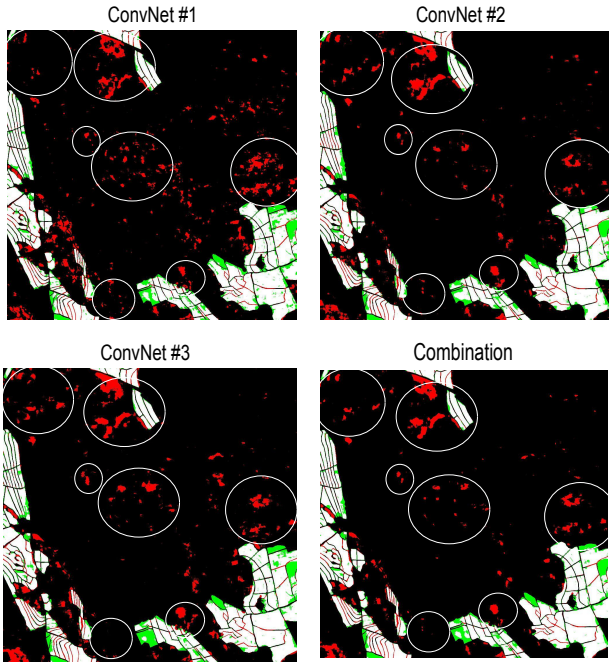

Fig. 7. Results for the combination and each single scale. Pixels correctly classified are shown in white (true positive) and black (true negative) while misclassified pixels are displayed in red (false positive) and green (false negative). We highlight some false positive group of pixels with white circles.

Fig. 8 illustrate an example of results comparing the proposed method against the HSMC-Boost baseline. One can observe that the main difference in these examples is that the proposed approach produces less false negative than the baseline. On the other hand, our approach produced more false positives.

Overall, our approach seems to be promising in reducing two main problems found in the baselines: (1) to discriminate recently planted coffee crops; and (2) to detect paths between the crops. As pointed by *dos Santos et al.* [6], most of the HSMSC-Boost classification errors are related to confusion caused by recently planted coffee crops, which usually appear in light blue in the composition of colors displayed. The proposed approach achieve good better in those areas. Moreover, it was more effective in assigning the class "non-coffee" to the paths between crops, as can be also observed in Fig. 8. The more the number of "black lines" between coffee crops the more accurate was the classification of paths.

The regions in red in Fig. 8(c) indicates most of the false positives produced by the proposed approach are due to dense native vegetation canopy. We believe the misclassified pixels can be better classified by including largest context windows

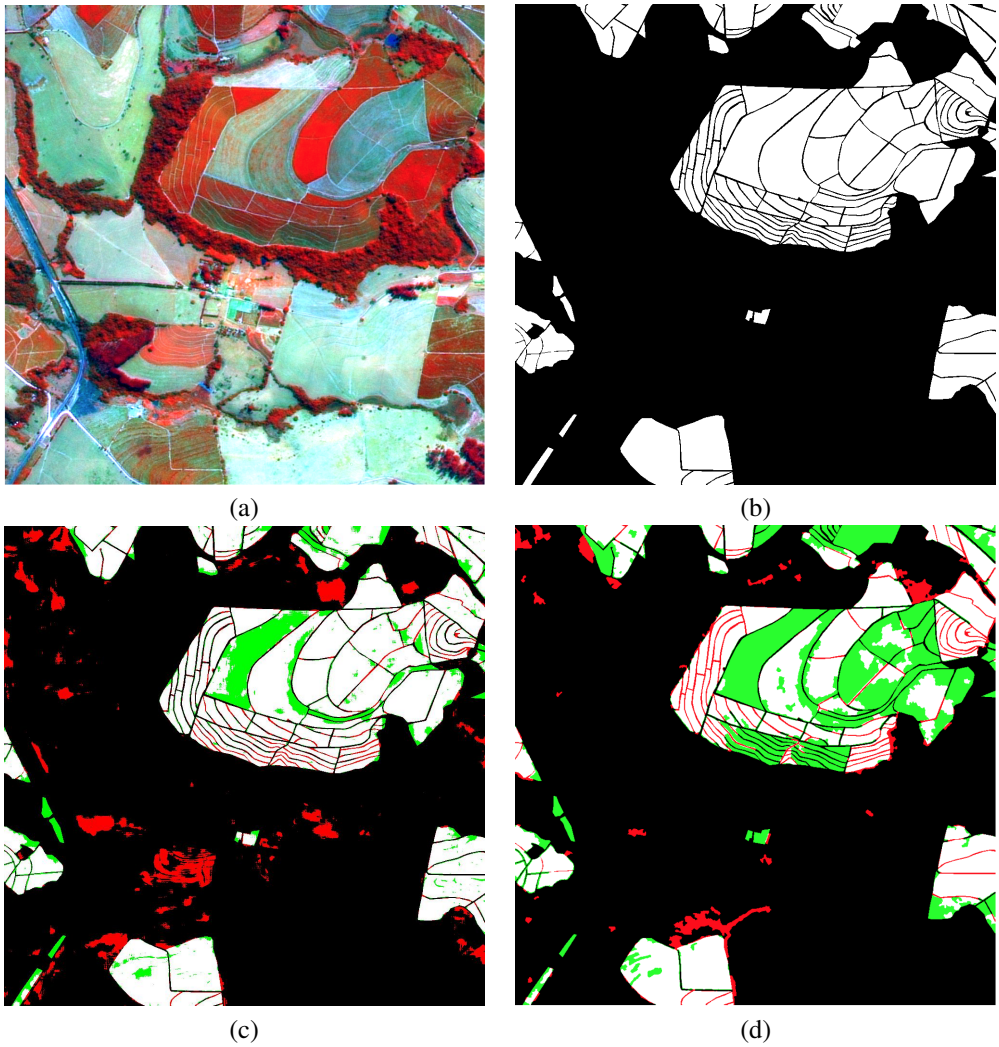

Fig. 8. Example of results: (a) input image, (b) ground truth, (c) the proposed approach, and (d) HMSC-Boost. Pixels correctly classified are shown in white (true positive) and black (true negative) while the errors are displayed in red (false positive) and green (false negative).

in the process. Also, these pixels are easier to remove by using some post-processing approaches than the misclassified regions produced by HSMC-Boost and other segmentation-based methods found in the literature. Note that the proposed approach misclassifies some very small group or even isolated pixels.

V. Conclusions and Future Work

In this paper, we propose a new approach based on Convolutional Neural Networks to learn spatial feature arrangements from remote sensing images in multiple scales aiming at the recognition of coffee crops. Experimental results show that the combination of ConvNets designed for semantically segment using different-size input really improves the final map in comparison with single scales. Our approach also achieves promising results in coffee crop recognition when compared to baselines.

As future work, we intend to apply the proposed approach in other applications. We also plan: to analyze larger context windows; to investigate more multi-scale fusion strategies; to develop a strategy to perform fine-tuning; and to address the imbalanced dataset training problem.

Acknowledgment

This work was partially financed by CNPq (grants 449638/2014-6, 307010/2014-7, 428333/2016-8), CAPES, and Fapemig (APQ-00768-14). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] M. Li, L. Ma, T. Blaschke, L. Cheng, and D. Tiede, "A systematic comparison of different object-based classification techniques using high spatial resolution imagery in agricultural environments," *International Journal of Applied Earth Observation and Geoinformation*, vol. 49, pp. 87–98, 2016.

[2] M. Vogels, S. de Jong, G. Sterk, and E. Addink, "Agricultural cropland mapping using black-and-white aerial photography, object-based image analysis and random forests," *International Journal of Applied Earth Observation and Geoinformation*, vol. 54, pp. 114 – 123, 2017.

[3] A. Chemura, O. Mutanga, and T. Dube, "Separability of coffee leaf rust infection levels with machine learning methods at sentinel-2 msi spectral resolutions," *Precision Agriculture*, pp. 1–23, 2016.

[4] J. A. dos Santos, F. A. Faria, R. T. Calumby, R. da S. Torres, and R. A. C. Lamparelli, "A genetic programming approach for coffee crop recognition," in *IEEE International Geoscience & Remote Sensing Symposium*, 2010.

[5] K. Nogueira, W. Schwartz, and J. A. dos Santos, "Coffee crop recognition using multi-scale convolutional neural networks," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2015, pp. 67–74.

[6] J. A. dos Santos, P. Gosselin, S. Philipp-Foliguet, R. da S. Torres, and A. X. Falcão, "Multiscale classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, pp. 3764–3775, 2012.

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

[8] K. Nogueira, W. O. Miranda, and J. A. dos Santos, "Improving spatial feature representation from aerial scenes by using convolutional networks," in *SIBGRAPI*, Aug 2015, pp. 289–296.

[9] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Learning to semantically segment high-resolution remote sensing images," in *International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 3566–3571.

[10] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, Feb 2017.

[11] J. Benediktsson, J. Chanussot, and W. Moon, "Advances in very-high-resolution remote sensing [scanning the issue]," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 566–569, March 2013.

[12] Y. Yang and S. Newsam, "Comparing sift descriptors and gabor texture features for classification of remote sensed imagery," in *International Conference on Image Processing*, 2008, pp. 1852–1855.

[13] J. A. dos Santos, O. A. B. Penatti, and R. da S. Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification," in *International Conference on Computer Vision Theory and Applications*, Angers, France, May 2010, pp. 203–208.

[14] R. Bouchiha and K. Besbes, "Comparison of local descriptors for automatic remote sensing image registration," *Signal, Image and Video Processing*, vol. 9, no. 2, pp. 463–469, 2013.

[15] J. dos Santos, O. Penatti, P. Gosselin, A. Falcao, S. Philipp-Foliguet, and R. Torres, "Efficient and effective hierarchical feature propagation," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. PP, no. 99, pp. 1–12, 2014.

[16] F. Faria, D. Pedronette, J. dos Santos, A. Rocha, and R. Torres, "Rank aggregation for pattern classifier selection in remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 4, pp. 1103–1115, April 2014.

[17] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 439–451, 2014.

[18] P. Tokarczyk, J. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 280–295, Jan 2015.

[19] A. Barsi and C. Heipke, "Artificial neural networks for the detection of road junctions in aerial images," *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. 34, no. 3/W8, pp. 113–118, 2003.

[20] O. Firat, G. Can, and F. Yarman Vural, "Representation learning for contextual object and region detection in remote sensing," in *International Conference on Pattern Recognition*, Aug 2014, pp. 3708–3713.

[21] C. Hung, Z. Xu, and S. Sukkarieh, "Feature learning based approach for weed classification using high resolution aerial images from a digital camera mounted on a uav," *Remote Sensing*, vol. 6, no. 12, pp. 12 037–12 054, 2014.

[22] H. Xie, S. Wang, K. Liu, S. Lin, and B. Hou, "Multilayer feature learning for polarimetric synthetic radar data classification," in *IEEE International Geoscience & Remote Sensing Symposium*, July 2014, pp. 2818–2821.

[23] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175–2184, April 2015.

[24] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised feature extraction of hyperspectral images," in *International Conference on Pattern Recognition*, 2014.

[25] M. E. Midhun, S. R. Nair, V. T. N. Prabhakar, and S. S. Kumar, "Deep model for classification of hyperspectral image using restricted boltzmann machine," in *International Conference on Interdisciplinary Advances in Applied Computing*, 2014, pp. 35:1–35:7.

[26] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2014.

[27] D. Tuia, R. Flamary, and N. Courty, "Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions," *{ISPRS} Journal of Photogrammetry and Remote Sensing*, no. 0, pp. –, 2015.

[28] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on*, 2015, pp. 44–51.

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

[30] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[31] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," in *Advances in Neural Information Processing Systems*, 2013, pp. 351–359.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[33] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, March 1977.