Coffee Crop Recognition Using Multi-scale Convolutional Neural Networks

Keiller Nogueira, William Robson Schwartz, and Jefersson A. dos Santos

Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{keiller.nogueira, william, jefersson}@dcc.ufmg.br

Abstract. Identifying crops from remote sensing images is a fundamental to know and monitor land-use. However, manual identification is expensive and maybe impracticable given the amount data. Automatic methods, although interesting, are highly dependent on the quality of extracted features, since encoding the spatial features in an efficient and robust fashion is the key to generating discriminatory models. Even though many visual descriptors have been proposed or successfully used to encode spatial features, in some cases, more specific description are needed. Deep learning has achieved very good results in some tasks, mainly boosted by the feature learning performed which allows the method to extract specific and adaptable visual features depending on the data In this paper, we propose two multi-scale methods, based on deep learning, to identify coffee crops. Specifically, we propose the Cascade Convolutional Neural Networks, or simply CCNN, that identifies crops considering a hierarchy of networks and, also, propose the Iterative Convolutional Neural Network, called ICNN, which feeds a same network with data several times. We conducted a systematic evaluation of the proposed algorithms using a remote sensing dataset. The experiments show that the proposed methods outperform the baseline consistent of state-of-the-art components by a factor that ranges from 3 to 6%, in terms of average accuracy.

Keywords: Deep Learning; Coffee Crop; Remote Sensing; Feature Learning;

1 Introduction

The use of Remote Sensing Images (RSIs) as a source of information is very common in several areas, such as agrobusiness. A lot of knowledge can be extracted from these images including geolocation of events (burned forest, for example), productivity forecast, and crop recognition. In this work, we focus on the latter task, specifically, we aim at identifying coffee crops in RSIs.

Considering this kind of plantation, the identification of crops is essential to know and monitor the land-use, helping to define new expansion strategies of the land or to estimate the feasible production amount. Although interesting,

recognizing coffee regions in RSIs is not a trivial task. First, because coffee usually grows in mountainous regions, which causes shadows and distortions in the spectral information. Second, the growing of coffee is not a seasonal activity, and, therefore, in the same region, there may be coffee plantations of different ages (high intraclass variance).

The identification process, which, in our case, can be described as locate and classify the crops, is an open problem in the pattern recognition field [5]. The most common strategy uses a combination of segmentation algorithms, visual features extraction techniques and machine learning methods. Some works [6] combine these steps with a multi-scale strategy, resulting in a more robust method. In all these cases, visual features are extracted from regions of a segmented image using some auxiliary method, such as low-level or mid-level one, and, then used with some machine learning approach. Although this method has been successfully applied to RSIs [5], some applications require more specific descriptors. In this way, the neural networks distinguish from other methods, since it can learn specific image features depending on the problem.

As introduced, in this paper, we are particularly interested in identifying coffee crops in RSIs. Therefore, we formulate this task by using a deep learning strategy, i.e., we propose **two** multi-scale methods using Convolutional Neural Network (CNN). First, we propose the **Cascade Convolutional Neural Network**, or simply CCNN, which is, in this case, composed of three network levels that process images with same dimension. Specifically, after every level, unclassified images are decomposed into smaller patches, which are resized into a predefined size and given as input to the subsequent level. The resize step changes the image composition allowing the networks to capture different features at each level. Second, we propose the **Iterative Convolutional Neural Network**, or just ICNN, which has only one neural network that processes the input data three times, being equivalent to the the CCNN method. Actually, after processing the data, unclassified patches are split and resized, going back again into the same network.

Moreover, we are concerned in design a method robust enough to handle real world data (even from different locations), so it can be a useful tool for any activity involving crops recognition around the globe. Thus, the proposed methods were designed and trained using real data of two entire counties, that have distinct image characteristics (mountains, etc). Specifically, the experiments were conducted using one county as training and the other as test.

In practice, we claim the following benefits and contributions over existing solutions: (i) Our main contribution is **two novel algorithms** capable of identify region of interest in real world RSIs using deep learning paradigm, and (ii) a systematic set of experiments, using real world data reveals that our algorithm improves upon a baseline composed of state-of-the-art components, by a factor that ranges from 3% to 6% in terms of average accuracy.

The paper is structured as follows. Related work is presented in Section 2. Section 3 presents the methodology. Experimental protocol as well as obtained

results are discussed in Section 4. Finally, in Section 5 we conclude the paper and point promising directions for future work.

2 Related Work

The development of algorithms for spatial extraction information is a hot research topic in the remote sensing community [2], which has been mainly boosted 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 [7], 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 [7]. Anyway, the general conclusion is that ordinary descriptors can achieve suitable results in most of applications, but not all. However, 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 combination of spatial descriptors [9], including several ones using multi-scale strategy [6, 7]. In these approaches, an essential step is extracting the feature at various segmentation scales, which could be expensive, depending on the strategy, since features would need to be extracted from each scale, for example.

However, even the combination of visual descriptors may not achieved satisfactory results and more robust features are needed. In this way, deep learning distinguish from other methods, since it can learn specific image features depending on the problem. Many works have been proposed to learn spatial feature descriptors [13]. Moreover, new effective hyperspectral and spatio-spectral feature descriptors [11] have been developed mainly boosted by the deep learning growth in recently years.

The proposed methods are very different from others in the literature. First, the proposed approach is capable of create a thematic map without any use of auxiliary methods. For the best of our knowledge, there is no other method capable of doing this. Second, as introduced, accuracy is highly dependent on the quality of extracted features. Thus, a method that learns adaptable and specific spatial features based on the images, such as the ones based on deep learning, could exploits better the feasible information available on the data. In this work, we experimentally demonstrate the robustness of our approach by achieving state-of-the-art results in a challenging dataset composed of high resolution remote sensing images.

3 Methodology

In this section, we present the proposed methods for identification of crops. The network architecture is presented first in Section 3.1 while the proposed methods are presented in Section 3.2.

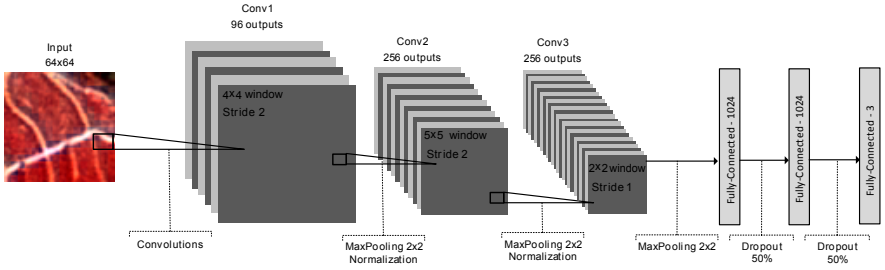

Fig. 1. The proposed Convolution Neural Network architecture with six layers.

3.1 Network Architecture

To achieve higher discrimination power with deep representations, the final network architecture, presented in Figure 1, is composed of six stacked layers: 3 convolutional (followed by max pooling and Local Response Normalization (LRN)), 2 fully-connected and a final classifier layer. All layers are composed of Rectifier Linear Units (ReLUs). Also, to prevent overfitting, the dropout method [10] was employed. At the end of the network, a softmax was used as classification layer. As mentioned, this architecture was used in both proposed methods, being the base of all the methodology employed in this work, which is presented next.

3.2 Multi-scale Convolutional Neural Network

The first multi-scale method proposed is the **Cascade Convolutional Neural Network model (CCNN)**, which is a hierarchical model composed of three levels[1], that always process tiles of 64×64 pixels, since this is the required input size of the proposed network. As mentioned, the same architecture was employed in all levels but with some differences related to the classification layer and the training data, depending on the level.

Considering the classification layer, in the first two levels, tiles must be classified into three possible classes. A threshold approach, based on the number of coffee pixels of the patch, was employed to select the class of each tile. Thus, a tile could be: (i) coffee, if a patch has, at least, 90% of coffee pixels, (ii) non-coffee, if a patch has, at maximum, 10% of coffee pixels, and mixed (or undefined), otherwise. How the last level must classify the remaining tiles, it has only two possible classes: coffee, patches with at least, 50% of coffee pixels, and non-coffee, otherwise. Considering all available training data, the first level network receive a small amount of patches while the last one is trained with a large amount of data, since between each level a tile is split and resized into a new patch, increasing the amount of available training data for the subsequent level.

Figure 2 presents a overview of the CCNN method. The first level network processes a small amount of tiles and, the ones classified into the mixed class are split into patches of 32×32 pixels, resized, and processed by the second level

[1] In this case, only three network levels were used based on a cost-benefit analysis.

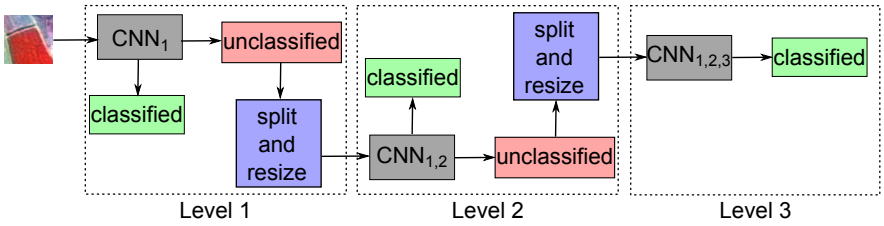

Fig. 2. A overview of the Cascade Convolutional Neural Network model. The subscript number of the convolutional symbolizes the quantity of data available for training each level.

network. Once more, unclassified tiles are again split into patches of 16×16 pixels and resized. The last level network is responsible to finally classify the remaining tiles. At the end, a class is associated to each tile and a new image may be recomposed, showing the regions of interest, in this case, the coffee crops.

The second method proposed is the **Iterative Convolutional Neural Network (ICNN)**, which has only one neural network that processes the input data three times, being equivalent to the CCNN method. Actually, after processing the data once, unclassified patches are split and resized, going back again into the same network. Just like the CCNN, this method uses the architecture proposed in Section 3.1, trained with all tiles split and resized into 64×64 pixels patches. These patches has three possible classes (coffee, non-coffee and mixed) independent of the iteration. The class of each tile were defined following the same protocol used in the first two levels of the CCNN method. However, by doing this, the last iteration, which must classify all remaining tiles into coffee or non-coffee classes, could classify tiles into a unwanted mixed class. A work around is to change the class of these tiles to the second class with higher probability. Thus, we force the last iteration to classify the remaining tiles, as intended.

A overview of the proposed method is presented in Figure 3a. The first iteration process tiles of 64×64 without resize. Unclassified tiles are split into patches of 32×32 pixels, resized and processed into the same network. The same occurs for the last iteration, which split the tiles into patches of 16×16 pixels, resized and processed, for the last time, into the same network.

4 Experimental Evaluation

In this section, we present the experimental setup and results.

4.1 Setup

Dataset. To evaluate the proposed methods, we used a multispectral high-resolution scene dataset, which is composed of **huge** scenes taken by the SPOT sensor in 2005 over two entire counties in the State of Minas Gerais, Brazil: Guaranésia and Guaxupé. Figure 3 shows some samples of these classes. As

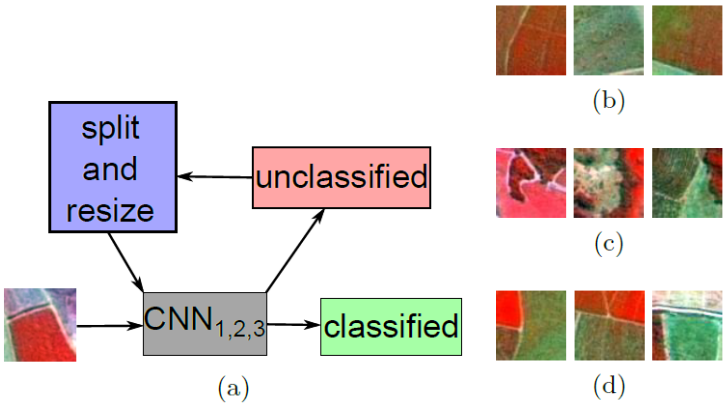

Fig. 3. (a): A overview of the Iterative Convolutional Neural Network model. (b)-(d): Respectively, coffee, non-coffee and mixed samples of the coffee dataset.

mentioned, this dataset was partitioned into tiles of 64×64, 32×32 and 16×16 pixels, generating, for Guaranésia, 21,600, 86,400 and 345,600 tiles, and, for Guaxupé, 17,280, 69,120 and 276,480 regions. Although interesting, this dataset has several challenges, such as: (i) high intraclass variance, caused by different crop management techniques, (ii) scenes with different plant ages, since coffee is an evergreen culture and, (iii) images with spectral distortions caused by shadows, since Minas Gerais is a mountainous region.

Baselines. As baseline, we consider the most common strategy that uses a combination of segmentation algorithms, visual features extraction techniques and machine learning methods. In this case, we have used SLIC [1], which has achieving good results for remote sensing images [12]. As visual features, BIC [4], which is the most suitable descriptor to describe coffee crops, as pointed out by [6], was employed. As machine learning technique, RBF-SVM was used. The CCNN paradigm was simulated in the baseline by extracting three different segmentation maps with different granularity.

Experimental Protocol. As introduced, we devised our experiments to evaluate the performance of the proposed methods considering a real world scenario. Thus, the protocol used consider one county for training and other for testing. Since there is much more non-coffee areas than coffee ones, the metric used to evaluate the proposed methods were the average accuracy, which is calculated by averaging the pixel accuracy for each class. The proposed networks were built using Caffe framework [8], since it is more suitable due to its simplicity and support to parallel programming using CUDA. Furthermore, all computational experiments presented were performed on a 64 bits Intel i7 4,960X machine with 3.6GHz of clock, 64 GB of RAM memory and GeForce GTX980. A drawback of deep learning strategy is the large number of parameters, which are, in this

case, five different ones: learning rate, weight decay, momentum, maximum iterations and step size (which defines the number of iterations until the learning is divided by a constant value (gamma) equals to 0.1). Select the best value for each parameter, as well as the best network architecture, is totally experimental, which requires a high number of tests and a well-structured protocol. In this case, the networks and its parameters were adjusted by considering a full set of experiments guided by [3]. After all the setup experiments, the best values for the learning rate, weight decay, step size, momentum and max iterations were 0.01, 0.001, 10,000, 0.9 and 30,000, respectively.

4.2 Results and Discussion

For the proposed methods, the processing time, for each county, took around one hour to be completed. At the end, the CCNN method yielded average accuracy around 57% and 63%, for Guaxupé and Guaranésia, respectively. Both results outperforms the baseline by a factor varying from 2 to 6%, in terms of average accuracy. The ICNN method also outperform the baseline, but was less effective than the CCNN approach. However, the baseline is more hand-working, since segments and features need to be extracted first to be, then, used with some machine learning technique, while the proposed methods learns all at once. Furthermore, it is worth to mention that agricultural scenes is very hard to classify since the method must to differentiate among different vegetation [6].

Method	Guaranésia	Guaxupé
SLIC+BIC+SVM-RBF	57.89	55.86
CCNN	**69.33**	**57.98**
ICNN	60.22	56.08

Table 1. Results, in terms of average accuracy (%), of the proposed methods and the baseline for the coffee dataset.

5 Conclusions and Future Work

In this paper, we propose two multi-scale methods based on Convolutional Neural Networks to identifying coffee crops from remote sensing images, considering a real world scenario. Experimental results show that the CCNN method is more effective and robust than all others, achieving state-of-the-art, in terms of average accuracy, for coffee crop identification, considering two entire counties. As future work, we intend to evaluate new datasets and applications. We also consider to use some different strategies, such as fine-tuning.

Acknowledgment

This work was partially financed by Brazilian National Research Council – CNPq (grant 449638/2014-6), the Coordination for the Improvement of Higher Educa-

tion Personnel – CAPES (DeepEyes Project), the Minas Gerais Research Foundation – FAPEMIG (grants APQ-00768-14 and APQ-00567-14). The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the GeForce GTX980 GPU used for this research.

References

[1] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 11 (2012), 2274–2282.

[2] Benediktsson, J., Chanussot, J., and Moon, W. Advances in very-high-resolution remote sensing [scanning the issue]. 566–569.

[3] Bengio, Y. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 437–478.

[4] de O. Stehling, R., Nascimento, M. A., and Falcao, A. X. A compact and efficient image retrieval approach based on border/interior pixel classification. In *International Conference on Information and Knowledge Management* (2002).

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

[6] dos Santos, J. A., Faria, F. A., da S Torres, R., Rocha, A., Gosselin, P.-H., Philipp-Foliguet, S., and Falcao, A. Descriptor correlation analysis for remote sensing image multi-scale classification. In *International Conference on Pattern Recognition* (Nov 2012), pp. 3078–3081.

[7] dos Santos, J. A., Penatti, O. A. B., Gosselin, P.-H., Falcao, A. X., Philipp-Foliguet, S., and da S Torres, R. Efficient and effective hierarchical feature propagation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7*, 12 (Dec 2014), 4632–4643.

[8] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* (2014).

[9] Schindler, K. An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing 50*, 11 (2012), 4534–4545.

[10] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research 15*, 1 (2014), 1929–1958.

[11] Tuia, D., Flamary, R., and Courty, N. Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions. *Journal of Photogrammetry and Remote Sensing*, 0 (2015).

[12] Vargas, J. E., Saito, P. T., Falcao, A. X., de Rezende, P. J., and dos Santos, J. A. Superpixel-based interactive classification of very high resolution images. In *SIBGRAPI Conference on Graphics, Patterns and Images* (2014).

[13] Zhang, F., Du, B., and Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing 53*, 4 (April 2015), 2175–2184.