https://doi.org/10.1109/CVPRW.2015.7301382

Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains?

Otávio A. B. Penatti
Advanced Technologies Group
SAMSUNG Research Institute
Campinas, SP, 13097-160, Brazil
o.penatti@samsung.com

Keiller Nogueira, Jefersson A. dos Santos
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, MG, 31270-010, Brazil
{keiller.nogueira,jefersson}@dcc.ufmg.br

Abstract

In this paper, we evaluate the generalization power of deep features (ConvNets) in two new scenarios: aerial and remote sensing image classification. We evaluate experimentally ConvNets trained for recognizing everyday objects for the classification of aerial and remote sensing images. ConvNets obtained the best results for aerial images, while for remote sensing, they performed well but were outperformed by low-level color descriptors, such as BIC. We also present a correlation analysis, showing the potential for combining/fusing different ConvNets with other descriptors or even for combining multiple ConvNets. A preliminary set of experiments fusing ConvNets obtains state-of-the-art results for the well-known UCMerced dataset.

1. Introduction

The recent impressive results of methods based on deep learning for computer vision applications brought fresh air to the research and industrial community. We could observe real improvements in several applications such as image classification, object and scene recognition, face recognition, image retrieval, and many others.

Deep learning for computer vision is usually associated with the learning of features using an architecture of connected layers and neural networks. They are usually called Convolutional (Neural) Networks or ConvNets. Before deep learning has attracted the attention of the community in the latest years, the most common feature descriptors were shallow without involving machine learning during feature extraction. Common visual descriptors, which are still interesting alternatives for feature extraction, are mid-level (bags of visual words – BoVW) and global low-level color and texture descriptors (e.g., GIST [24], color histograms, and BIC [9]). BoVW descriptors are somewhat a step in the direction of feature learning, as the visual codebook is usually learned for dataset of interest. Global descriptors, however, have a pre-defined algorithm for extracting the image feature vector, independently of the dataset to be processed. They tend to be less precise, but they are usually fast to compute.

ConvNets have shown astounding results even in datasets with different characteristics from which they were trained, feeding the theory that deep features are able to generalize from one dataset to another. Successful ConvNets freely available in the literature are OverFeat and Caffe, which were originally trained to recognize the 1,000 object categories of ImageNet [28, 18]. OverFeat, for instance, has already shown that it works remarkably well in applications like flower categorization, human attribute detection, bird sub-categorization, and scene retrieval. In [28], Razavian et al. suggest that features obtained from deep learning should be the primary candidate in most visual recognition tasks.

The use of deep learning for remote sensing is rapidly growing. A considerable number of works appeared very recently proposing deep strategies for spatial and spectral feature learning. Even though, to the best of our knowledge, there is still no evaluation of pre-trained ConvNets in the aerial and remote sensing domain. Therefore, this paper adds two more domains in which pre-trained ConvNets, like OverFeat and Caffe, are evaluated and compared with existing image descriptors.

In this paper, besides evaluating ConvNets in a different domain, we also perform an evaluation of several other image descriptors, including simple low-level descriptors and mid-level representations. The evaluation is based on the classification of aerial image scenes and on remote sensing images aiming at differentiating coffee and non-coffee crop tiles. We also conduct a correlation analysis in order to identify the most promising descriptors for selection/fusion. The correlation analysis includes even different ConvNets.

We can summarize the main contributions of this paper as follows:

- evaluation of the generalization power of ConvNets from everyday objects to the aerial and remote sensing domain,
- comparative evaluation of global descriptors, BoVW descriptors, and ConvNets,
- correlation analysis among different ConvNets and among different descriptors.

On top of that, we performed preliminary experiments for fusing ConvNets and obtained state-of-the-art results for the classification of aerial images using the UCMerced dataset. For the remote sensing domain, we created a new dataset, which is publicly released.

The remainder of this paper is organized as follows. Section 2 presents related work. The ConvNets and the descriptors evaluated in this paper are presented in Section 3. The experimental setup and datasets are presented in Section 4. In Section 5, we present and discuss the results obtained. Finally, Section 6 concludes the paper.

2. Related Work

As far as the remote sensing (RS) community, motivated by the accessibility to high spatial resolution data, started using more than pixel information for classification, the study of algorithms for spatial extraction information has been a hot research topic. Although many descriptors have been proposed or successfully used for RS image processing [42, 11, 3], some applications require 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 [12]. Anyway, the general conclusion is that ordinary descriptors can achieve suitable results in most of applications. However, higher accuracy rates are yielded by the combination of complementary descriptors that exploits late fusion learning techniques. In this context, frameworks have been proposed for selection of spatial descriptors in order to learn the best algorithms for each application [10, 7, 14, 35]. In [10], the authors analyzed the effectiveness and the correlation of different low-level descriptors in multiple segmentation scales. They also proposed a methodology to select a subset of complementary descriptors for combination. In [14], Faria et al. proposed a new method for selecting descriptors and pattern classifiers based on rank aggregation approaches. Cheriyadat [7] proposed a feature learning strategy based on Sparse Coding. The strategy learns features in well-known datasets from the literature and uses for detection of buildings in larger image sets. Tokarczyk et al. [35] proposed a boosting-based approach for the selection of low-level features for very-high resolution semantic classification.

Artificial Neural Networks have been used for RS classification for a long time [2]. But, similarly to the computer vision community, its massive use is recent and chiefly 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 [15, 17, 45]. Firat et al. [15] proposed a method based on ConvNets for object detection in high-resolution remote sensing images. Hung et al. [17] applied ConvNets to learn features and detect invasive weed. Zhang et al. [45] proposed a deep feature learning strategy that exploits a pre-processing salience filtering. In [41], the authors presented an approach to learn features in Synthetic Aperture Radar (SAR) images. Moreover, the "deep learning boom" has been seen as the golden opportunity for developing effective hyperspectral and spatio-spectral feature descriptors [29, 23, 6, 36].

In the computer vision community, with the release of pre-trained ConvNets, like OverFeat [31] and Caffe [18], they started being evaluated in different applications than the ones they were trained for. In [28], for instance, a ConvNet trained for recognizing 1,000 object categories has shown very good results even in applications like bird subcategorization, scene retrieval, human attribute detection and others, which are considerably different than everyday object recognition. Those facts raised the issue about the generality of the features computer by ConvNets.

In this paper, we go in this direction of evaluating pre-trained ConvNets in different domains. It is worth to mention that, to the best of our knowledge, there is no other work in literature that evaluate the feasibility of using deep features from general computer vision datasets in remote sensing applications. In addition, no other work in the literature has evaluated the complementarity of deep features aiming at fusion or classifier ensemble.

3. Feature Descriptors

In this section, we describe the ConvNets, low-level (global), and mid-level (BoVW) descriptors we have used. The descriptors we have selected for evaluation were mainly based on previous works [42, 11, 37, 26, 44, 12], in which they were evaluated for remote sensing image classification, texture and color image retrieval/classification, and web image retrieval. Besides the evaluation of ConvNets, we also selected a set of other types of descriptors. Our selection includes simple global descriptors, like descriptors based on color histograms and variations, and also descriptors based on bags of visual words (BoVW).

3.1. Convolutional Networks

In this section, we provide details about the ConvNets used in this work, which are OverFeat [31] and Caffe [18].

OverFeat [31] is a deep learning framework focused on ConvNets. It is implemented in C++ and trained with the Torch7 package[1]. OverFeat was trained on the ImageNet 2012 training set [30], which has 1.2 million images and 1,000 classes, and it can be used to extract features and/or to classify images.

OverFeat has two models available. A small (*fast* – OverFeat$_S$) and a larger one (*accurate* – OverFeat$_L$), both having similarities with the network of Krizhevsky et al. [19]. The main differences to the Krizhevsky's network are: (i) no response normalization and (ii) non-overlapping pooling regions. OverFeat$_L$ differs in some details, including: (i) one more convolutional layer and, (ii) the number and size of feature maps, since different number of kernels and stride were used for the convolutional and the pooling layers. OverFeat$_S$ is more similar to the Krizhevsky's network, differing only in the number and size of feature maps.

The main differences between the two OverFeat networks are the stride of the first convolution, the number of stages and the number of feature maps [31]. When using the feature extractor of OverFeat, we obtain a feature vector of 4,096 dimensions, which directly corresponds to the output of layers 19 for OverFeat$_S$ and 22 for OverFeat$_L$. They are the last fully-connected layer, which is composed, in both networks, of 4,096 kernels (one dimension per kernel).

Caffe or Convolutional Architecture for Fast Feature Embedding [18] is a fully open-source framework that affords clear and easy implementations of deep architectures. It is implemented using C++ with support to CUDA®, a NVidia® parallel programming based on graphics processing units (GPU). Caffe uses Protocol Buffer language, which makes it easier to create new architectures. It also has several other functionalities, such as fine-tuning strategy, layer visualization and feature extraction. Just like OverFeat, we are interested in extracting features using a pre-trained network, which is, in this case, almost the same of Krizhevsky's network [19], proposed for the ILSVRC 2012 competition [30], with two differences: (i) during the training no data-augmentation was used (to increase the number of training examples) and, (ii) the order of pooling and normalization was switched, since Caffe does pooling before normalization. The pre-trained model of Caffe was obtained using the same dataset of the competition and basically the same parameters of the Krizhevsky's network [19]. Therefore, Caffe was also trained to recognize 1,000 categories of everyday objects. This framework allows feature extraction for any layer of the network. In our experiments, features from the last layer were extracted, which results in a vector of 4,096 features, one for each kernel of the last fully-connected layer, just like OverFeat.

According to [31], the number of parameters for each ConvNet are (in millions): 60, 145, and 144 (Krizhevsky, OverFeat$_S$, and OverFeat$_L$, respectively). The number of connections are (in millions): 2,810 and 5,369 (OverFeat$_S$ and OverFeat$_L$, respectively).

[1]Torch is a scientific computing framework with wide support for machine learning algorithms freely available at http://www.torch.ch.

3.2. Low-Level descriptors

The Color Autocorrelogram (**ACC**) descriptor [16] is a color descriptor which aims to encode spatial color distribution in the image. Such information is extracted by computing the probabilities of having two pixels of color c_i in a distance d from each other. ACC performed well in previous works for natural image representation [26].

Border-Interior Pixel Classification (**BIC**) [9] is a simple color descriptor which computes two color histograms for an image: one for border pixels and other for interior pixels. BIC obtained good results in previous works for web image retrieval [26] and for remote sensing image classification [11].

Local Color Histogram (**LCH**) [33] computes a color histogram for an image divided into tiles. An independent histogram is computed for each tile and then, they are concatenated to form the image feature vector.

Statistical Analysis of Structural Information (**SASI**) [5] is based on a set of sliding windows, which are covered in different ways. SASI was very effective for texture discrimination in previous works [26].

Local Activity Spectrum (**LAS**) [34] captures the spatial activity of a texture in the horizontal, vertical, diagonal, and anti-diagonal directions separately. According to [26], LAS achieved good results for texture discrimination in terms of both effectiveness and efficiency.

GIST [24] provides a global holistic description representing the dominant spatial structure of a scene. GIST is popularly used for scene representation [13].

Histogram of Oriented Gradients (**HOG**) [8] computes histograms of gradient orientations in each position of a sliding window. We used HOG in different configurations, varying the cell size in 20×20, 40×40 and 80×80 pixels, but keeping the orientation binning in 9 bins.

3.3. Mid-Level descriptors

Bags of visual words (**BoVW**) and their variations [32, 39, 4, 21, 25, 1, 27] are mid-level representation based on a codebook of visual discriminating patches (visual words). They compute statistics about the visual word occurrences in the image. BoVW descriptors have been the state of the art for several years in the Computer Vision community and are still important candidates to perform well in many tasks.

We used BoVW in several different configurations: sparse sampling (Harris-Laplace detector) or dense sampling (grid of circles with 6 pixels of radius); SIFT and OpponentSIFT as descriptors [37]; visual codebooks of sizes 100, 1000, 5000, and 10000; hard or soft assignment (with

σ = 90 or 150); and average, max pooling or WSA pooling [25].

To differentiate them in the experiments, we used the following naming: BX^{w}_{cp}, where X is S (sparse sampling) or D (dense sampling); w is the codebook size; c refers to the coding scheme used, h (hard), s (soft); p refers to the pooling technique used, a (average), m (max), or W (WSA).

4. Experimental Setup

The main objective of this paper is to evaluate the generalization capacity of ConvNets. The ConvNets used here were trained to recognize 1,000 object categories and we evaluate their generality in an experimental scenario with aerial and remote sensing image scenes. We also included other image descriptors aiming at contrasting their effectiveness with ConvNets and also aiming at verifying if they can provide complementary results. The experiments are conducted in a classification protocol, in which the dataset is split into training and testing sets and image feature vectors from the training set are used to feed a machine learning classifier. The test set is then used for evaluating the learned classifiers in terms of classification accuracy.

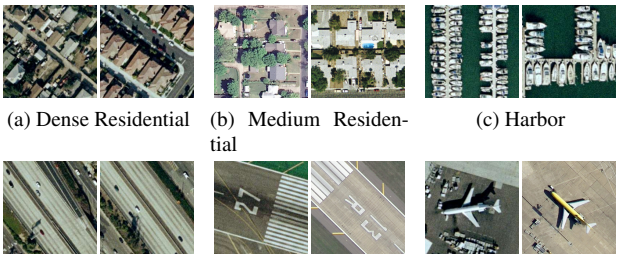
4.1. Datasets

We have chosen datasets with different properties in order to better evaluate the descriptors robustness and effectiveness. The first one is a multi-class land-use dataset that contains aerial high resolution scenes in the visible spectrum. The second dataset has multispectral high-resolution scenes of coffee crops and non-coffee areas.

4.1.1 UCMerced Land-use

This dataset [43] is composed of 2,100 aerial scene images with 256×256 pixels divided into 21 land-use classes selected from the United States Geological Survey (USGS) National Map. Some class samples are shown in Figure 1. These images were obtained from different US locations for providing diversity to the dataset. The categories are: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. The dataset presents highly overlapping classes such as "dense residential", "medium residential" and "sparse residential", which mainly differ in the density of structures. It is publicly available [43].

In the experiments, as OverFeat and Caffe have special requirements in the resolution of input images (231×231 for OverFeat$_S$, 221×221 for OverFeat$_L$, and 227×227 for Caffe), we decided to resize all the images to 231×231 pixels discarding the aspect ratio. Caffe implicitly crops the center of each image to obtain the needed resolution.



(a) Dense Residential (b) Medium Residential (c) Harbor

(d) Freeway (e) Runway (f) Airplane

Figure 1: Examples of the UCMerced Land Use Dataset.

The resized images were used also for all the descriptors evaluated.

4.1.2 Brazilian Coffee Scenes

This dataset, publicly released with this paper[2], is a composition of scenes taken by the SPOT sensor in 2005 over four counties in the State of Minas Gerais, Brazil: Arceburgo, Guaranésia, Guaxupé, and Monte Santo. This is a very challenging dataset since there are many intraclass variance caused by different crop management techniques. Also, coffee is an evergreen culture and the South of Minas Gerais is a mountainous region, which means that this dataset includes scenes with different plant ages and/or with spectral distortions and shadows.

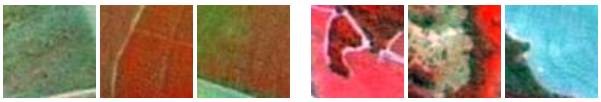
The whole image set of each county was partitioned into multiple tiles of 64×64 pixels. For this dataset, it was considered only the green, red, and near-infrared bands, which are the most useful and representative ones for discriminating vegetation areas. The identification of coffee crops (i.e., ground-truth annotation) was performed manually by agricultural researchers. The creation of the dataset is performed as follows: tiles with at least *85%* of coffee pixels were assigned to the *coffee* class; tiles with less than *10%* of coffee pixels were assigned to the *non-coffee* class; the remaining tiles were categorized as "undefined" or "mixed" and discarded in our analysis. In summary, considering the tiles of all the four counties, the dataset has 36,577 tiles of non-coffee, 1,438 tiles of coffee, and 12,989 mixed tiles.

In the experiments, images were resized to 231×231 pixels only for OverFeat and Caffe, given their requirements on the size of the input images. For all the other descriptors, images were kept in the original resolution of 64×64 pixels.

4.2. Experimental protocol

We carried out experiments for both datasets with a 5-fold cross-validation protocol using a linear SVM as classifier. Concerning the UCMerced dataset, each of the 5 folds has around 420 images and is unbalanced in terms of

[2]The Brazilian Coffee Scenes dataset as well as the folds used in this paper are available for download at: www.patreo.dcc.ufmg.br/downloads/brazilian-coffee-dataset/



(a) Coffee (b) Non-coffee

Figure 2: Example of coffee and non-coffee samples in the Brazilian Coffee Scenes dataset. The similarity among samples of opposite classes is notorious. The intraclass variance is also perceptive.

the number of samples per class. For the Brazilian Coffee Scenes dataset, 4 folds have 600 images each and the 5^{th} has 476 images, all folds are balanced with coffee and non-coffee samples (50% each). Therefore, the 5 folds comprise all the 1,438 coffee tiles explained in Section 4.1.2, but not all the non-coffee tiles.

We report results in terms of average accuracy and standard deviation among the 5 folds. For a given fold, we compute the accuracy for each class and then compute the average accuracy among the classes. This accuracy is used to compute the final average accuracy among the 5 folds.

After evaluating each descriptor in both datasets independently, we have also performed a correlation analysis. It is based on the correlation coefficient ρ defined in [20]:

$$\rho_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11}+N^{10})(N^{01}+N^{00})(N^{11}+N^{01})(N^{10}+N^{00})}} \quad (1)$$

where N^{11} and N^{00} represent the percentage of cases when both descriptors predicted correctly or incorrectly, respectively; while N^{10} and N^{01} represent cases of divergence, the first descriptor correctly predicts the class but not the second, and vice-versa, respectively; i and k refers to descriptor i and descriptor k.

ρ varies from -1 to 1. The smaller the value, the least correlated the descriptors and, therefore, the most promising to be combined. The opposite, i.e., higher values, mean that descriptors are more correlated. The best scenario for combining descriptors is the one that both descriptors present high accuracy values and are few correlated.

After evaluating the descriptors correlation, we also present some preliminary experiments for fusing feature vectors of ConvNets.

Descriptor implementations The implementations of ACC, BIC, LCH, SASI, and LAS descriptors follow the specifications of [26]. GIST implementation is the one used in [13] with the parameters discussed therein.[3] We used the HOG implementation of VLFeat [40]. The low-level feature extraction of BoVW descriptors was based on the implementation of van de Sande et al. [38]. For BoVW, in UCMerced dataset we used SIFT [22] to describe each patch, but in Coffee dataset, we used OpponentSIFT [37], as color should provide more discriminating power. The feature vectors of OverFeat are exactly the 4,096 values corresponding to the outputs of the small and large networks. The same for Caffe.

[3] http://lear.inrialpes.fr/software (as of March 14th, 2015).

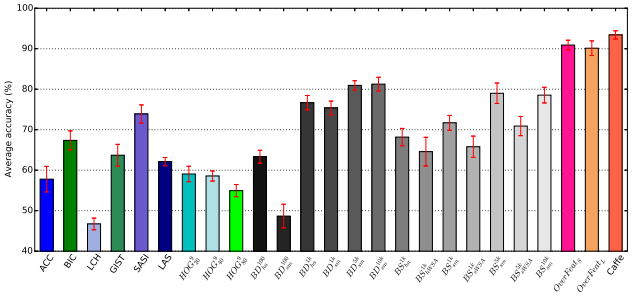


Figure 3: Average accuracy of the descriptors for the UCMerced Land-use Dataset. ConvNets achieve the highest accuracy rates.

5. Results and Discussion

In this section, we present the experimental analysis, which are divided in three phases: effectiveness evaluation, correlation analysis, and feature fusion.

5.1. Effectiveness Evaluation

In Figure 3, we show the average accuracy of each descriptor for the UCMerced dataset. We can observe that ConvNets features achieve the highest accuracy rates ($\geq$ 90%). Caffe features have a higher average accuracy (93.42% $\pm$ 1.00) in comparison to OverFeat (90.91 $\pm$ 1.19 for the small and 90.13 $\pm$ 1.81 for the large network). SASI is the best global descriptor (73.88% $\pm$ 2.25), while the best BoVW configurations are based on dense sampling, 5 or 10 thousand visual words and soft assignment with max pooling ($\sim$81%). The results illustrate the capacity of ConvNets generalize to the aerial domain.

In Figure 4, we show the average accuracy of each descriptor for the Coffee dataset. We can see that Caffe and OverFeat features achieve high accuracies, with Caffe having better results (84.82% $\pm$ 0.97). However, global color descriptors like BIC and ACC achieve high accuracies, with BIC outperforming all the other descriptors (87.03% $\pm$ 1.17). The BIC algorithm for classifying pixels in border or interior basically separates the images into homogeneous and textured regions. Then, a color histogram is computer for each type of pixel. As for the Coffee dataset the differences between classes may be not only in texture but also in color properties, BIC could encode well such differences.

The best BoVW configurations are again based on dense sampling, 5 or 10 thousand visual words and soft assignment with max pooling. They have comparable results to OverFeat.

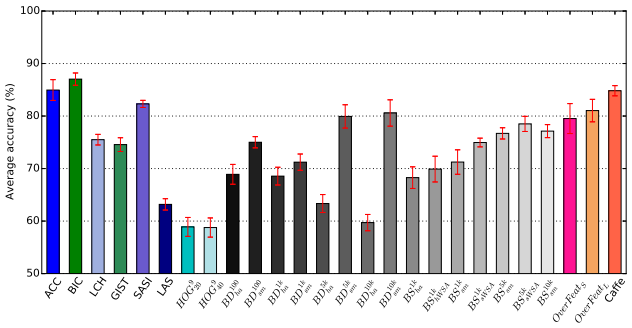


Figure 4: Average accuracy of the descriptors for the Brazilian Coffee Scenes dataset. We can note that, although the ConvNets perform well, some global color descriptors, like BIC, achieve higher accuracies.

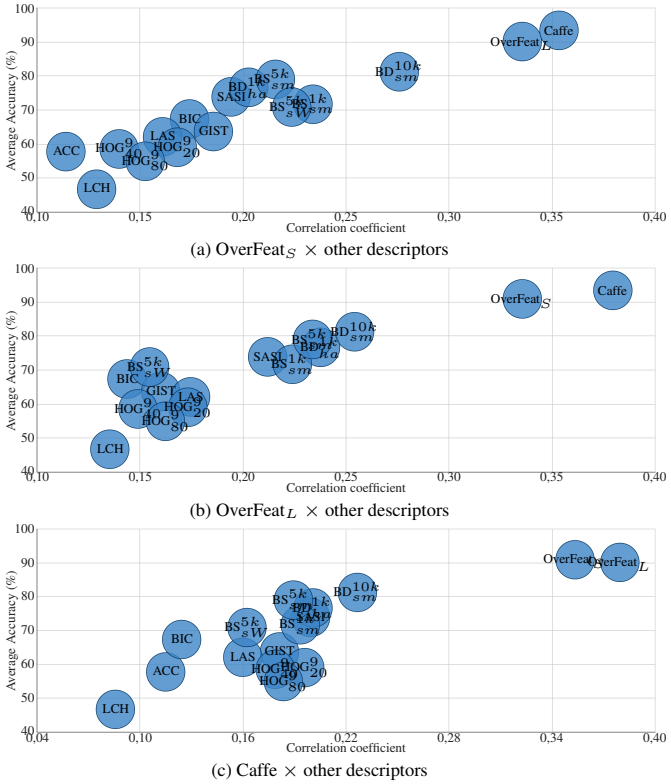
Although ConvNets were not the best descriptors for the Coffee dataset, they could still perform well. This is interesting specially because the ConvNets used here were trained to recognize objects, which is a very different scenario in relation to recognizing coffee regions. This also shows the generalization power of ConvNets.

A possible reason for the ConvNets to perform better in aerial dataset than in the agricultural one is due to the particular intrinsic properties of each dataset. The UCMerced have more complex scenes, composed of a lot of small objects (e.g., buildings, cars, airplanes). Many of these objects are composed of similar visual patterns in comparison with the ones found in the dataset used to train the ConvNets, with salient edges and borders. Concerning the Brazilian Coffee Scenes dataset, it is composed of finer and more homogeneous textures where the patterns are much more overlapping visually and more different than everyday objects. The color/spectral properties are also important in this dataset, which fit with results reported in other works [12, 14].

5.2. Correlation analysis

We carried out a correlation analysis with the results of all the descriptors in each dataset. Figures 5 and 6 show the correlation coefficient versus the average accuracy for each ConvNet in relation to all the other descriptors. For better visualization, some descriptors are not shown in the figures. The top-left located descriptors in each graph should be more promising to combine to the ConvNets, as they are least correlated to them and have higher accuracy.

One can observe that ConvNets tend to agree (higher correlation values) with the descriptors that have better accuracies. This is somewhat expected given Equation 1. Therefore, for UCMerced dataset (Figures 5), we can see that the ConvNets are usually more correlated to other ConvNets. However, we can notice that the correlation values are not high, being closer to zero than to 1. This fact indicates a good potential for fusing even different ConvNets. In Figure 6, ConvNets are more correlated to BIC, ACC and the other ConvNets.



(a) OverFeat$_S$ × other descriptors

(b) OverFeat$_L$ × other descriptors

(c) Caffe × other descriptors

Figure 5: Correlation analysis of ConvNets versus all the other descriptors in the UCMerced dataset. We can note that the ConvNets are more correlated to each other and less correlated to the other descriptors. Even though, the correlation values are not high even between ConvNets.

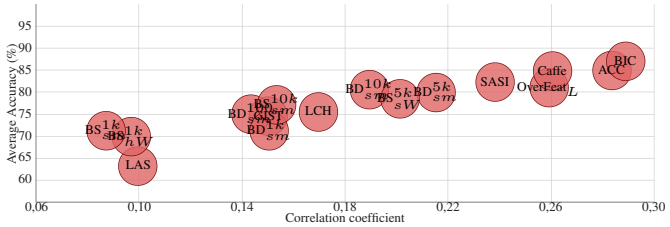
5.3. Feature fusion

We have also conducted preliminary experiments of feature fusion. Our intention was to verify the results when combining multiple ConvNets. Our fusion strategy is a simple concatenation of the feature vectors computed by each ConvNet, without even performing any normalization step. The results were verified by a fold-by-fold paired test with confidence level of 95%.

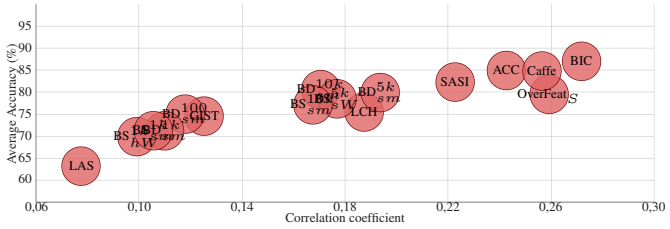
We can see in Table 1 that for the UCMerced dataset, the combination of OverFeat and Caffe performed remarkably well, achieving almost 100% of average accuracy. To the best of our knowledge, these are state-of-the-art results for this dataset. For the coffee dataset, improvement was observed only when combining the two OverFeat networks.

6. Conclusions

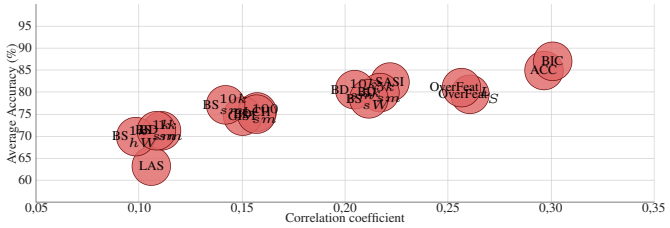
In this paper, we experimentally evaluated ConvNets trained for recognizing everyday objects in the aerial and re-



(a) OverFeat$_S$ × other descriptors



(b) OverFeat$_L$ × other descriptors



(c) Caffe × other descriptors

Figure 6: Correlation analysis of ConvNets versus all the other descriptors in the Brazilian Coffee Scenes dataset. ConvNets are more correlated to BIC, ACC, and the other ConvNets, which perform better in this dataset.

mote sensing image classification task. This evaluation adds two more scenarios in which pre-trained ConvNets generalize well. The results show that **yes**, the ConvNets generalize well even in domains considerably different from the ones they were trained for. We also compared ConvNets with a set of other visual descriptors. ConvNets achieved the highest accuracy rates for aerial images, while for coffee scenes they perform well but do not outperform low-level color descriptors, such as BIC. A correlation analysis of ConvNets with other ConvNets as well as with other visual descriptors shows that they can potentially be combined to achieve even better results. Preliminary experiments fusing two ConvNets obtain state-of-the-art results for the well-known UCMerced dataset.

As future work, we intend to explore more opportunities for fusing ConvNet features and other descriptors based on our correlation analysis presented in this paper. We also would like to evaluate the impact of using features from lower layers of the ConvNets as well as to train ConvNets with specific remote sensing data. The evaluation of ConvNets in other remote sensing and aerial datasets is another interesting future work.

Table 1: Results when fusing ConvNets.

(a) UCMerced dataset

Descriptors	**Avg. Accuracy**	**Improves?**
OverFeat$_S$ + OverFeat$_L$	93.05 ± 0.88	Yes
OverFeat$_S$ + Caffe	$\mathbf{99.36 \pm 0.63}$	Yes
OverFeat$_L$ + Caffe	$\mathbf{99.43 \pm 0.27}$	Yes

(b) Brazilian Coffee Scenes dataset

Descriptors	**Avg. Accuracy**	**Improves?**
OverFeat$_S$ + OverFeat$_L$	83.04 ± 2.00	Yes
OverFeat$_S$ + Caffe	79.01 ± 1.53	No
OverFeat$_L$ + Caffe	79.15 ± 1.70	No

Acknowledgments

This work was partially financed by CNPq (grant 449638/2014-6), CAPES, and Fapemig (APQ-00768-14). We thank the RECOD lab of Institute of Computing, University of Campinas, Brazil, for the infrastructure for running experiments. We also thank Rubens Lamparelli and Cooxupé for the image sets.

References

[1] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo. Pooling in image representation: the visual codeword point of view. *CVIU*, 117(5):453–465, 2013. 3

[2] A. Barsi and C. Heipke. Artificial neural networks for the detection of road junctions in aerial images. *Int. Arch. of Photogrammetry Remote Sensing and Spatial Inf. Sciences*, 34(3/W8):113–118, 2003. 2

[3] R. Bouchiha and K. Besbes. Comparison of local descriptors for automatic remote sensing image registration. *Signal, Image and Video Processing*, 9(2):463–469, 2013. 2

[4] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, pages 2559–2566, 2010. 3

[5] A. Çarkacioglu and F. Yarman-Vural. Sasi: a generic texture descriptor for image retrieval. *Pattern Recognition*, 36(11):2615 – 2633, 2003. 3

[6] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu. Deep learning-based classification of hyperspectral data. *IEEE JSTARS*, 7(6):2094–2107, 2014. 2

[7] A. M. Cheriyadat. Unsupervised feature learning for aerial scene classification. *IEEE TGRS*, 52(1):439–451, 2014. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 3

[9] R. de O. Stehling, M. A. Nascimento, and A. X. Falcao. A compact and efficient image retrieval approach based on border/interior pixel classification. In *CIKM*, pages 102–109, 2002. 1, 3

[10] J. A. dos Santos, F. A. Faria, R. da S Torres, A. Rocha, P.-H. Gosselin, S. Philipp-Foliguet, and A. Falcao. Descriptor correlation analysis for remote sensing image multi-scale classification. In *ICPR*, pages 3078–3081, 2012. 2

[11] 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 *VISAPP*, pages 203–208, 2010. 2, 3

[12] J. A. dos Santos, O. A. B. Penatti, P.-H. Gosselin, A. X. Falcao, S. Philipp-Foliguet, and R. da S Torres. Efficient and effective hierarchical feature propagation. *IEEE JSTARS*, 7(12):4632–4643, 2014. 2, 6

[13] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *ACM CIVR*, pages 19:1–19:8, 2009. 3, 5

[14] F. Faria, D. Pedronette, J. dos Santos, A. Rocha, and R. Torres. Rank aggregation for pattern classifier selection in remote sensing images. *IEEE JSTARS*, 7(4):1103–1115, 2014. 2, 6

[15] O. Firat, G. Can, and F. Yarman Vural. Representation learning for contextual object and region detection in remote sensing. In *ICPR*, pages 3708–3713, 2014. 2

[16] J. Huang, S. R. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *CVPR*, pages 762–768, 1997. 3

[17] 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*, 6(12):12037–12054, 2014. 2

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 1, 2, 3

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012. 3

[20] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003. 5

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006. 3

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 5

[23] 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 *ICONIAAC*, pages 35:1–35:7, 2014. 2

[24] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 1, 3

[25] O. A. B. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, and R. da S. Torres. Visual word spatial arrangement for image retrieval and classification. *Pattern Recognition*, 47(2):705–720, 2014. 3, 4

[26] O. A. B. Penatti, E. Valle, and R. da S. Torres. Comparative study of global color and texture descriptors for web image retrieval. *JVCIR*, 23(2):359–380, 2012. 2, 3, 5

[27] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *ECCV*, pages 143–156, 2010. 3

[28] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *CVPR Workshop*, pages 512–519, 2014. 1, 2

[29] A. Romero, C. Gatta, and G. Camps-Valls. Unsupervised deep feature extraction of hyperspectral images. In *ICPR*, 2014. 2

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014. 3

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229v4*, 2014. 2, 3

[32] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003. 3

[33] M. J. Swain and D. H. Ballard. Color indexing. *IJCV*, 7(1):11–32, 1991. 3

[34] B. Tao and B. W. Dickinson. Texture recognition and image retrieval using gradient indexing. *JVCIR*, 11(3):327–342, 2000. 3

[35] P. Tokarczyk, J. Wegner, S. Walk, and K. Schindler. Features, color spaces, and boosting: New insights on semantic classification of remote sensing images. *IEEE TGRS*, 53(1):280–295, 2015. 2

[36] D. Tuia, R. Flamary, and N. Courty. Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions. *J. of Photogrammetry and Remote Sensing*, (0):–, 2015. 2

[37] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9):1582–1596, 2010. 2, 3, 5

[38] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Empowering visual categorization with the gpu. *Trans. on Multimedia*, 13(1):60–70, 2011. 5

[39] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *TPAMI*, 32:1271–1283, 2010. 3

[40] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008. 5

[41] H. Xie, S. Wang, K. Liu, S. Lin, and B. Hou. Multilayer feature learning for polarimetric synthetic radar data classification. In *IGARSS*, pages 2818–2821, 2014. 2

[42] Y. Yang and S. Newsam. Comparing sift descriptors and gabor texture features for classification of remote sensed imagery. In *ICIP*, pages 1852–1855, 2008. 2

[43] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *ACM GIS*, pages 270–279, 2010. 4

[44] Y. Yang and S. Newsam. Geographic image retrieval using local invariant features. *IEEE TGRS*, 51(2):818–832, 2013. 2

[45] F. Zhang, B. Du, and L. Zhang. Saliency-guided unsupervised feature learning for scene classification. *IEEE TGRS*, 53(4):2175–2184, 2015. 2