Could Multimedia approaches help Remote Sensing Analysis?

Amina Ben Hamida, Alexandre Benoit, Patrick Lambert, Chokri Ben-Amar

Could Multimedia approaches help Remote Sensing Analysis?

Amina Ben Hamida[*,**], *Alexandre Benoit*[*], *Patrick Lambert*[*], *Chokri Ben Amar*[**]

[*] LISTIC, Université Savoie Mont Blanc
74940 Annecy le Vieux, France
{amina.ben-hamida,alexandre.benoit, patrick.lambert}@univ-smb.fr

[**] REGIM, Ecole Nationale d'Ingénieurs de
Sfax Route de Soukra, B. P. W, 3038 Sfax,
Tunisie chokri.benamar@ieee.org

Abstract—

The paper explores how multimedia approaches used in image understanding tasks could be adapted and used in remote sensing image analysis. Two approaches are investigated: the classical Bag of Visual Words (BoVW) approach and the Deep Learning approach. Tests are performed for the classification of the UC Merced Land Use Dataset which provide better results than the state of the art.

Keywords—Bag of visual words, Deep Learning, Remote sensing, classification

I. INTRODUCTION

Remote sensing is a rapidly growing field that has been witnessing various breakthroughs over the last decades. In fact, the advent of satellites such as the recently launched Sentinels is allowing a high acquisition rate coupled with a very rich images' content which resulted in the data deluge era. Given the overwhelming masses of available remote sensing data, a review of the currently used approaches is needed. One solution is to resort to the multimedia field and select the most suitable candidates among the existing methods. The inspiration can be derived from the fact of not tackling the image as a unity which can be achieved whether by the use of more locally focused descriptors or by the increase of its semantic level. Migrating from the color, texture, shape… descriptors [1] to richer ones such as SIFT, SURF, FREAKS LBP… is one effective solution. Therefrom, comes the opportunity to resort to approaches with higher semantic level such as the Bag of Visual Words (BoVW), or to step into the data mining, machine learning or deep learning fields.

In the remote sensing literature, we can find a few recent attempts which transpose these approaches, mostly dedicated to the multimedia domain, for remote sensing image analysis. As it was first successfully introduced by Yang [2] who defines a BoVW-based approach for the classification of land areas that are reshaped by humans (parking, golf courses, etc.) or later introduced by Dos Santos to extract features independently from their shape through the use of BoVW for cultivated areas [3]. In [4], Vaduva proposes a deep learning algorithm for semantic annotation of very high resolution remote sensing images.

In this paper, we use the "UC Merced Land Use Dataset" and Yang [2] results as a baseline to explore the interest of the Bag of Visual Words and the Deep Learning approaches in the classification of remote sensing images. This dataset is a small one with low resolution images, but it works as a first step to demonstrate the potential of such approaches in the RS context.

The outline of this paper is as follows. In Section II, we carry out and test several improvements of the BoVW approach proposed by Yang [2]. In Section III, we adapt the deep learning architecture proposed in [10] and test it on the same dataset. In Section IV, we conclude with a short summary.

II. IMPROVEMENT OF BOVW APPROACH

A. The Bag of visual words

The "bag of visual words" approach [5] basically consists of: (1) extracting a large collection of local invariant features from a large set of images, (2) quantizing the features, often using k means clustering, in order to create a k sized vocabulary. Once the visual words are selected as the centers of the k clusters, (3) features of a new image can be translated into visual words by deciding which clusters they are nearest to and the new image is then mapped over an empirical space, captured into an histogram that counts how many times each word occurs within it. Supervised classification algorithms are applied afterwards, generally using SVM classifier, to find the link between different BoVW histograms and different semantic concepts.

For the sake of a good image representation, the features must be extracted from interesting locations or key points that are characterized by invariance toward geometric transformations including rotations, translations….

The literature encompasses many feature extraction techniques. However, we choose to work with both the Scale Invariant Feature Transform (SIFT) [6] and the Speed Up Robust Features (SURF) [7] as a baseline for further image analysis since they have been proven to be the best performers in many image retrieval approaches.

B. The Extentions of BoVW

We investigate and compare BoVW state of the art methods. Starting from the usual sparse, gray level SIFT detector, we experimented different options in terms of density, color space and spatial structure.

- *Dense feature extraction*

Contrarily to the basic sparse feature extraction, where the descriptors are only dedicated to the points of interest, the dense sampling consists in breaking the image into a number of patches and lately creates a descriptor for each selected patch center. Such strategy was first initialized by [8]. It allows a full coverage of the entire scene thanks to the equal contribution of all regions drawing the image. Thus, we end up with a constant number of extracted features allowing an accurate, yet costly representation of the objects. As it is deeply tackled in [9], although the dense sampling does not consider keypoint orientation and imposes scale, it succeeded to improve over sparse keypoint detection for BoVW classification.

- *BoVW applied to RGB images*

Usually the BoVW is applied to gray level images, in total absence of the color components. The idea is then to examine the contribution of the Red, Green and Blue channels in the process of the "visual vocabulary" creation.

In order to procure a BoVW for an RGB image, we need to first separate the channels. The features are then computed for each and every one of the three components independently. In that case, we acquire three 128 dimensioned SIFT descriptors for each keypoint that are concatenated into a 128×3= 384 long vector that feeds the BoVW toolchain. Note that color space change can be applied before descriptors extraction. For example, opponent color space and HSV are often used in the multimedia context.

C. Experiments

We have resorted to the "UC Merced Land Use Dataset[1]" [2] that was manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the country. The pixel resolution of this public domain imagery is 1 foot. 100 images measuring 256×256 pixels were manually selected for each of the following 21 classes: agricultural, airplane, baseball diamond, beach, buildings... Figure 1 shows a few images.

This dataset is of moderate size but has the advantage to allow us to compare to a previous study for a fair comparison. It also allows various techniques to be experimented and directions to be drawn for further analysis on much larger datasets.

Airplane Golf course Storage tank Mobile home park

Fig. 1. examples of the UC Merced Land Use Dataset

The process was made with 1000 sized vocabulary and a classification phase was performed using the Support Vector Machine (SVM) with and 'RBF' kernel.
The evaluation process consists in training all the methods on a specific subset of the dataset. It is made of 95% of the images with similar class proportions. Performances evaluation is carried on the remaining validation subset. 6 fold cross validation is used and we present average performance values. Each classification rate was calculated based on the number of correctly labeled images among all the collection of testing, and the presented results are an average rate over all classes.

As shown in table 1, the use of SIFT as a feature extraction tool ensures relatively high classification rates. Our implementation of the sparse features approach is equivalent to the one of Yang [2] and presents a similar performance baseline. However, the interference of the dense sampling instead of the classical sparse SIFT keypoints detection, significantly enhances the results (9% above the Yang baseline 78%). Some classes witness a better classification rate than others. For example, Chaparral, harbor, and parking perform much better than storage tanks and tennis classes.

Taking the color components into account, in the BoVW process, doesn't seem to increase the overall classification rate at first glance. But, going through all classes one by one and comparing their correctly labeled rate each one aside proves that such approach is highly dependent from the images scenes nature. Therefore, it's obvious that for some sort of classes such as 'mobile home park', 'golf course' and 'overpass', the RGB channels enhances the classification rate while decreases it for some other classes: 'airplane' and 'storage tanks'. Here again the combination of the color space and the dense sampling allows an improvement of performances.

TABLE 1. CLASSIFICATION RATE FOR BOVW AND ITS EXTENSIONS USING SIFT

	BoVW using SIFT				
	Yang sparse features baseline [2]	***Our sparse feature detection***	***Dense sampling***	***Sparse feature detection + RGB***	***RGB Dense sampling + RGB***
Classification rate	78%	79%	87%	78%	86%

[1] http://vision.ucmerced.edu/datasets/landuse.html

The results for the use of SURF as a feature extraction tool in this context are summarized in table 2. In fact, SURF here has proven to be outperforming the SIFT on different levels: whether it's handled with sparse or dense sampling or even with the RGB channels concept adoption.
It is important to note, that although SURF has better results, it still follows the same pattern as the SIFT performances toward the different BoVW extensions.

TABLE 2. CLASSIFICATION RATE FOR BOVW AND ITS EXTENSIONS USING SURF

	BoVW using SURF			
	Sparse feature detection	***Dense sampling***	***BoVW + RGB***	***RGB + Dense***
Classification rate	86%	91%	85%	90%

III. DEEP LEARNING APPROACH

A. CNN networks

The fast paste of technologies development and the continuous enhancement of images resolution stimulate an urgent need for images deeper analysis. More and more details and concepts are included in a single image; therefore, especially in the satellite imaging domain, the analysis at different scales and regions is now mandatory for a comprehensive, accurate description of the Data. For each concept to detect, a model should be provided according to the type of the targeted information, namely, pixels, objects, structures…. However, current BOW approaches are difficult to design in this aim and relying on a single hand crafted feature such as SIFT or SURF cannot be the perfect method for all the cases. As an illustration, SIFT is built to describe detailed areas and it won't be efficient to describe untextured areas such as lakes, deserts, etc.

A very different approach recently emerged and tries to answer such challenge by learning representations from the data. In this family of methods, Deep Learning with convolutional neural networks recently proved to progressively outperform all the other approaches in the multimedia community starting from the ILSVRC2012 challenge with the work proposed by Hinton&al [10]. Such methods inherited from the work of Yann LeCun [11] rely on neural layers that can be trained in a supervised manner using error back-propagation. It is now adaptable to a wide variety of context thanks to new operators, the availability of very large annotated training databases and massive computing machines such as Graphical Processing Units (GPU).

B. CNN components

We briefly describe the main components of a deep neural network and more details can be obtained from the following reference paper [10]. Basically, there's no unique architecture for the neural networks. Although it involves the same fundamental elements:

- Convolution layers (Conv) take as input only limited areas of the input space thus preserving locality. They correspond to a set of convolutionnal operations applied all over the input. Such layers, applied successively allow the semantic description of the image to increase along layers from edges, corners, to parts and objects.
- Fully Connected (FC) layers are similar to CL but are the classical neural net layer that takes as input all the available outputs of the previous layer thus ignoring locality information. It generally handles the majority of the parameters to train and is generally used at the end of the architecture. It finalizes the representation learning process by modeling the target concepts from a composition of their already high semantic level input features.
- Non linearities (NL) are a key property of neurons. They are thus systematically applied at the output of Conv and FC layers. They are historically Sigmoids but the recently introduced Rectified Linear Units (ReLU, f(x)=max(0,x)) proved more efficient network training.
- Pooling operators (Pool) which are used to subsample input data by applying fusion operators over local neighborhood such as max or average. It enforces robustness against translations.
- Local Contrast Normalization operators (LCN) that enhance local contrast and help luminance and contrast invariance.
- Depending on the task, a final layer may be used to format output to the required type. In a classification task, a Softmax layer is generally used to generate a probability distribution and predict a single class over a set of mutually exclusive classes. SoftMax can also be used as a loss in the training stage. Such loss measures the error made by the net when misclassifying samples. It is used to compute layer weights corrections (gradients) in a chain-rule manner (back-propagation) from net top to bottom.

C. CNN architecture

Several network designs have already been proposed and a critical point concerns its depth that should be enough to get high semantic level and transferable enough learned representations. The counterpart is the computational cost. A compact yet efficient architecture is AlexNet [10] but deeper ones generally improve as discussed and compared in [12]. In this paper, we want to study the learned multimedia features representation of CNN to our context. To do so, we propose to use a AlexNet like network with weights learned from the multimedia dataset of ILSVRC2012 and we show that it can adapt to remote sensing images with few changes. Chosen network implementation is CaffeNet from the caffe library [13] with the available pretrained weights.

As illustrated in figure 3, CaffeNet is composed of 5 Conv layers, all using ReLU non linearities. Some of them are

followed by Pooling and Contrast normalization steps. The net ends with 3 fully connected layers each being followed by a ReLU non linearity. The last layer has an output dimension of size 1000 which corresponds to the 1000 imagenet multimedia classes on which the net has been trained for and is followed by a Softmax layer that generates class probability distributions.
We briefly present the considerations taken into account when choosing the mostly performing model. Therefore, first, we have selected 5 conv layers based on the complexity level and the lack of the amount of the available data. Secondly, since the pooling layers allow a significant decrease on the spatial level, the introduction of 3 layers is a decision made based on the hardware capacities. Finally, the fully connected layers number is fixed to 3 in order to overcome the over fitting problems resulting from the increased model capacity.

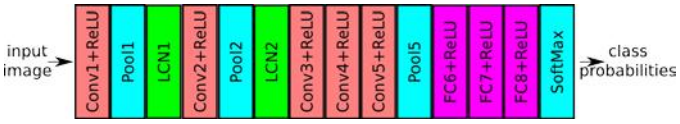

Fig. 2. CaffeNet deep neural network architecture

Then, we use this network in the two following configurations.

CaffeNet_FineTune: we follow the classical fine tuning procedure: we keep the original net with its initial weights but replace the last layer by a new one having uninitialized weights and a number of outputs corresponding to the new number of classes of the new classification task. In a training stage, only this last layer is retrained using a softmax loss.

CaffeNet_ fc7_features: we use the net as a feature extractor. To do so, we keep the entire original net but avoid the last specialized layer (FC8+SoftMax) and only use the net to extract the feature vectors from FC7 layer that feed a linear SVM classifier.

D. Experiments

Using the same training and testing datasets design described with the BoVW evaluation, we report the obtained results in table 3. All images are resized to 227*227 pixels to comply with the neural net configuration.

First, the impressive performance of both approaches are surprising since all the layers of the net except the new FC8 have been trained on multimedia concepts recognition (animals, persons, sports, etc.) that are totally different from the ones that we try to classify here. Second, it appears that the linear SVM classification of features extracted from FC7 layer is slightly more efficient than training a new specialized neural layer. However, both approaches significantly outperform the hand crafted feature methods that we evaluated.

TABLE 3. Classification rate for Deep learning approaches

	Deep learning using original CaffeNet trained architecture	
	CaffeNet_FineTune	CaffeNet_fc7_features +Linear SVM
Classification rate	94%	95%

A per class analysis of the *CaffeNet_ fc7_features* approach shows that the few misclassifications made concern some resembling classes such as “dense residential” and “mobile homepark”, or “forest” and “agricultural”.

IV. CONCLUSION

We describe and compare the BoVW and its extensions to the deep Learning approaches in the use of remote sensing images classification field. Although the BoVW seems to be a robust tool for such purposes, deep learning has managed to take the lead. This work catalyzes more efforts toward the exploration of new enhancement paths and basically the validation of this work using a larger and more delicate database.
Another direction would be to train a deep architecture from scratch on a remote sensing database instead of using weights pretrained in other contexts but the availability of a large collection of annotated regions will be part of the challenge.

References

[1] James B. Campbell, Randolph H. Wynne. Introduction to Remote Sensing. *Guilford Press*, 2011, ISBN 160918176X, 978160918176

[2] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. *In* Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems *(GIS '10)*, ACM, New York, NY, USA, 270-279. DOI=10.1145/1869790.1869829.

[3] J.A. dos Santos, O. A. B. Penatti, R. da S. Torres, P.H. Gosselin, S. Philipp-Foliguet and A.X. Falcao, Improving Texture Description in Remote Sensing Image Multi-Scale Classification Tasks By Using Visual Words. In IAPR International Conference on Pattern Recognition. Tsukuba Science City, Japan, November 2012.

[4] C. Vaduva, I. Gavat, M. Datcu. Deep learning in very high resolution remote sensing image information mining communication concept. *in Signal Processing Conference (EUSIPCO)*, 2012 Proceedings of the 20th European , vol., no., pp.2506-2510, 27-31 Aug. 2012.

[5] G.Csurka, C. R. Dance, L. Fan, J.Willamowski and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision*, ECCV, pages 1–22, 2004.

[6] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision, vol. 60, no. 2, pages 91–110, November 2004

[7] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool. Speeded-Up Robust Features (SURF). Comput. Vis. Image Underst., vol. 110, no. 3, pages 346–359, June 2008.

[8] T. Tuytelaars. Dense interest points. In CVPR, pages 2281–2288. *IEEE*, 2010.

[9] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In: *Proceedings of the British machine vision conference (BMVC)*, 2011.

[10] A. Krizhevsky, I. Sutskever, G.E. Hinton ImageNet Classification with Deep Convolutional Neural Networks, NIPS : Neural Information Processing Systems, Lake Tahoe, Nevada, 2012.

[11] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, 86(11):2278-2324, November 1998,

[12] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the Devil in the Details: Delving Deep into Convolutional Networks, British Machine Vision Conference, 2014

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev J. Long,R. Girshick, S. Guadarrama, T. Darrell, Trevor, Caffe: Convolutional Architecture for Fast Feature Embedding,In *Proceedings of the ACM International Conference on Multimedia* (MM '14). ACM, New York, NY, USA, 675-678.