

CONTEXTUAL SUPERPIXEL DESCRIPTION FOR REMOTE SENSING IMAGE CLASSIFICATION

J. E. Vargas¹, A. X. Falcão¹, J. A. dos Santos², J. C. D. M. Esquerdo³, A. C. Coutinho³, J. F. G. Antunes³

¹Institute of Computing - University of Campinas, Campinas, Brazil

²Department of Computer Science - Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

³Embrapa Informática Agropecuária, Campinas, Brazil

ABSTRACT

In superpixel-based remote sensing image classification and any other supervised learning problem two principal factors define the success of a determined classifier: the descriptors and the selected training samples. While many works have been devoted to create accurate prediction models and to select informative training samples few works have focused on the extraction of descriptors for superpixels of remote sensing images. This work presents a scheme for superpixel description, based on the Bag of visual Words model, that include contextual information using data of adjacent superpixels. Experiments performed in two remote sensing images show a remarkable advantage of contextual descriptors against to widely used superpixel descriptors.

Index Terms— superpixel description, contextual information, bag of visual words

1. INTRODUCTION

High spatial resolution images are usually segmented into small regions, called *superpixels*, for image annotation. Color and texture descriptors are assigned to each superpixel and a subset of superpixels is used to train a classifier by using supervised learning. The classifier should then be able to annotate the remaining image regions. The success of the automatic annotation, however, depends fundamentally on the quality of the superpixel descriptors [1] and the training samples. We address the first problem by proposing a scheme to create *contextual superpixel descriptors* based on Bag of visual Words (BoW). It is well known in remote sensing that information from nearby pixels (context) improves image classification [2]. We extend this result for superpixels using BoW — a higher-level data representation which has caught increasing attention in remote sensing image classification [3, 4]. Contextual information is widely used in the classification phase [5, 2]. Different from those works, we use context to create better superpixel descriptors. Color and texture descriptors have been largely used and few works have been devoted to create more effective

descriptors for region-based remote sensing image classification that employ BoW for superpixel-based remote sensing image classification. In this work, we propose a contextual superpixel descriptor based on the BoW model, that consider the information of the superpixels itself and a determined set of neighbor superpixels. The classification experiments using two remote sensing images show a remarkable advantage of contextual descriptors against to widely used region descriptors. Furthermore, we compare our results with the superpixel descriptor proposed in [6], that is based on BoW and SIFT. We used Simple Linear Iterative Clustering (SLIC) algorithm [7] to generate superpixels from images. In [7], many state-of-the-art techniques are compared under two metrics: boundary recall and under-segmentation error. SLIC was found to be more effective according to both metrics.

The organization of the paper is as follows: In Section 2, we present the proposed scheme for contextual superpixel description. Section 3 describes the dataset we used for the experiments. Section 4 shows the descriptors performance comparing our method with several baseline descriptors. Finally, conclusions are drawn in Section 5.

2. CONTEXTUAL SUPERPIXEL DESCRIPTION USING BOW

The original idea in Bag of visual Words (BoW) is to extract features from local patches centered at given points in each image of a dataset, select a set of them, randomly or through clustering, to conform the codebook (set of code words or visual words), and then assign to each image the corresponding histogram of visual words as image descriptor. In our case, the dataset consists of superpixels from a given image. Therefore, each superpixel generates a histogram of code words by counting the code words of sampling points that fall inside the superpixel. Contextual information is then added by aggregating the histograms of code words from nearby superpixels, as detailed next.

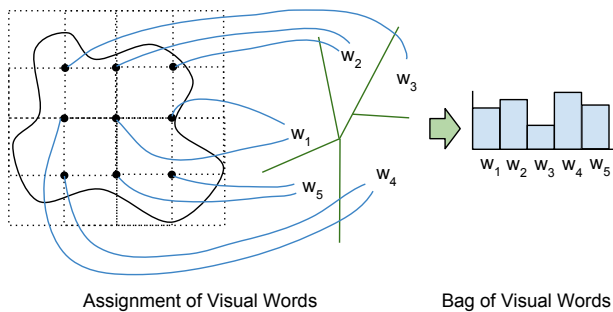


Fig. 1. Superpixel descriptor based on bag of visual words using dense grid sampling.

2.1. Superpixel description using BoW

A dense grid sampling (using every other pixel) is applied to select a set of pixels from the image. Color/texture features are then extracted from each 5×5 local patch around the selected pixels. A set of c pixels from the grid is randomly selected to conform the codebook and the closest code word is assigned to each pixel of the grid. The image is then segmented into superpixels and a histogram of code words is computed for each superpixel by considering the grid pixels inside it (Figure 1). The codebook size $c = 64$ was small because higher values were ineffective. We randomly select the codebook because it has the same results as clustering.

2.2. Contextual superpixel description

The oversegmented image is interpreted as a graph $G(V, E)$, where V is the set of superpixels and E contains the pairs of adjacent superpixels. A superpixel s_i is adjacent to a superpixel s_j if and only if a pixel from s_i is 4-connected to a pixel from s_j . We define a contextual t -neighborhood descriptor $D_i^t = \sum_{P(s_i, s_j)=t} D_j^0$ for superpixel s_i as the histogram created by aggregating histograms D_j^0 from superpixels s_j , where $P(s_i, s_j)$ is the shortest-path length between s_i and s_j in $G(V, E)$. In order to incorporate different levels of contextual information, we have also concatenated D_i^t for different values of t (Figure 2). For example, a contextual descriptor D_i of superpixel s_i can be created by the concatenation of the descriptors D_i^0 , D_i^1 and D_i^2 . We have also summarized many neighborhood levels into one. For instance, in the experiments, we have used:

$$D_i = (D_i^0, D_i^1, \sum_{r \in \{2,3\}} D_i^r), \quad (1)$$

where $\sum_{r \in \{2,3\}} D_i^r$ aggregate the 2- and 3-neighborhood descriptors into one component of D_i .

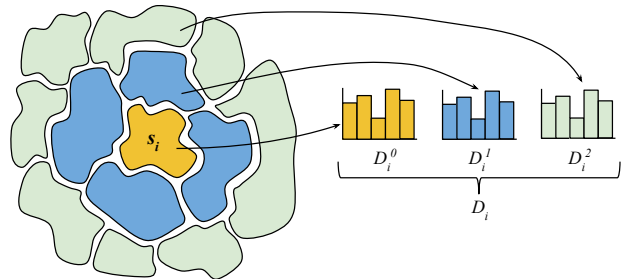


Fig. 2. Unnormalized contextual descriptor D_i of the superpixel s_i composed by concatenating: The unnormalized descriptor D_i^0 of s_i and its neighborhood descriptors D_i^1 and D_i^2 .

Including many contextual levels in the final descriptor may decrease the classification performance, because adjacent superpixels with different labels have similar contextual information. The separation of D_i^0 in an individual component of D_i helps to discriminate between two adjacent superpixels with different labels.

A similar contextual superpixel descriptor was proposed in [6] for object localization. It can be considered as a special case of the proposed one. They simply merge the descriptors of the superpixels that are less than t nodes away from the superpixel s_i in order to define its descriptor D_i . For $t = 3$, for instance,

$$D_i = \sum_{j \leq 3} D_i^j. \quad (2)$$

Given that the descriptors represent different numbers of local patches, we need to normalize every feature dividing it by the number of local patches. Finally, the new descriptor is given by the concatenation of the normalized histograms.

3. DATASETS

The experiments involved two datasets: CBERS and ROME. The CBERS dataset consists of superpixels from a multispectral image, as obtained by the CBERS-2 sensor (China-Brazil Earth-Resources Satellite) in 2007 over the state of Mato Grosso do Sul, Brazil. This image was annotated by agricultural researchers into 10 classes: Pasture, Vegetation/Forest, Farms, Annual Agriculture, Reforestation, Sugarcane, Vegetation/Savannah, Water, Urban area, and Occupied Floodplain. The annotation was performed using the color composition 3R, 4G, 2B, mapping the visible red spectrum into red, the near-infrared into green, and the visible green spectrum into blue. The ROME dataset consists of superpixels from a QuickBird image taken in 2004 over the Vatican City. It was annotated into 7 classes: Road, Tree, Shadow, Water, Building, Grass, and Soil.

4. EXPERIMENTS

We compared the performance of superpixel descriptors by randomly dividing the datasets 10 times into 10% of samples for training a SVM classifier and 90% of them for testing it. As baselines, we used low-level superpixel descriptors selected from [1]: Quantized Compound Change Histogram (QCCH), Local Binary Patterns (LBP), Color Histogram (CH), Border/Interior Pixel Classification (BIC), Mean Color (MC), and Color Autocorrelogram (ACC). For CH, BIC, and ACC we used histograms of 512 bins. The proposed descriptors used the BoW model, denoted by *Contextual BOW* (CBOW) the descriptors created by using Equation 1. CBOW-MC and CBOW-BIC are contextual descriptors created by computing MC and BIC over the local patches, respectively. The concatenation of these descriptors (CBOW-MC + CBOW-BIC) is denoted by *Concatenated Contextual BOW* (CCBOW). In order to also compare a BOW contextual descriptor with a contextual descriptor using only low-level features, we included *Contextual MC* (CMC) and *Contextual BIC* (CBIC) in the experiments, which are computed by Equation 1 using MC and BIC, respectively. The method NBOW-SIFT (NBOW — *Neighborhood BOW* using SIFT features on local patches)[6] is a particular case of ours, according to Equation 2. It may be interpreted as a baseline that uses BOW contextual information. We also used MC and BIC rather than SIFT over local patches for comparison. This descriptor is named CNBOW=(NBOW-MC + NBOW-BIC) (*Concatenated Neighborhood BOW*).

The SVM classifier used Gaussian kernel and its parameters were found by grid searching with 5-fold cross-validation in the training set. The comparison among descriptors used two classification metrics: Kappa index (κ) and Overall Accuracy.

Tables 1 and 2 show the Kappa index and overall accuracy obtained in the CBERS and ROME datasets, respectively, using the SVM classifier. The best performance is obtained by the proposed method CCBOW. Among the low-level descriptors, the ACC descriptor obtain the best performance in the CBERS and ROME datasets. The SIFT-based approach NBOW-SIFT proposed in [6] has the worst results among all contextual descriptors analyzed. The contextual descriptors CMC and CBIC obtain better results than MC and BIC, respectively. Figure 3 shows the classification results in a subset of the CBERS image using CCBOW and ACC (the best low-level descriptor).

Additionally, we assess the performance of ACC and CCBOW using the k-NN classifier, this is shown in Tables 3 and 4 for the CBERS and ROME datasets, respectively. As one can observe, the descriptor CCBOW with k-NN as classifier achieve better results than ACC using SVM. This shows that extracting good descriptors can be even more important than using a more sophisticated classifier.

Descriptors	CBERS			
	Kappa (κ)		Overall accuracy (%)	
	Mean	SD	Mean	SD
QCCH	0.1750	0.0054	49.90	1.58
LBP	0.2561	0.0235	55.77	0.85
CH	0.5860	0.0037	74.21	0.27
BIC	0.5909	0.0065	74.48	0.49
MC	0.5847	0.0066	74.08	0.68
ACC	0.6105	0.0035	75.76	0.22
CBOW-MC	0.6795	0.0021	79.82	0.13
CBOW-BIC	0.6474	0.0056	77.92	0.29
CCBOW	0.6952	0.0051	80.78	0.30
CMC	0.6249	0.0025	76.61	0.10
CBIC	0.6838	0.0029	80.16	0.18
NBOW-SIFT	0.4896	0.0275	69.36	0.98
CNBOW	0.6517	0.0091	78.51	0.33

Table 1. Region descriptors classification performance for the CBERS dataset using SVM.

Descriptors	ROME			
	Kappa (κ)		Overall accuracy (%)	
	Mean	SD	Mean	SD
QCCH	0.3805	0.0105	60.74	1.28
LBP	0.5234	0.0062	68.20	0.28
CH	0.7760	0.0039	84.70	0.22
BIC	0.7801	0.0047	85.01	0.28
MC	0.7637	0.0044	83.81	0.30
ACC	0.7857	0.0034	85.36	0.21
CBOW-MC	0.8074	0.0042	86.81	0.28
CBOW-BIC	0.8076	0.0065	86.79	0.40
CCBOW	0.8200	0.0027	87.60	0.17
CMC	0.7740	0.0040	84.45	0.22
CBIC	0.7992	0.0058	86.25	0.35
NBOW-SIFT	0.2299	0.0271	54.90	2.14
CNBOW	0.3041	0.0116	59.25	0.28

Table 2. Region descriptors classification performance for the ROME dataset using SVM.

Descriptors	CBERS			
	Kappa (κ)		Overall accuracy (%)	
	Mean	SD	Mean	SD
ACC	0.5993	0.0036	75.00	0.31
CCBOW	0.6364	0.0035	77.02	0.25

Table 3. Region descriptors classification performance for the CBERS dataset using k-NN.

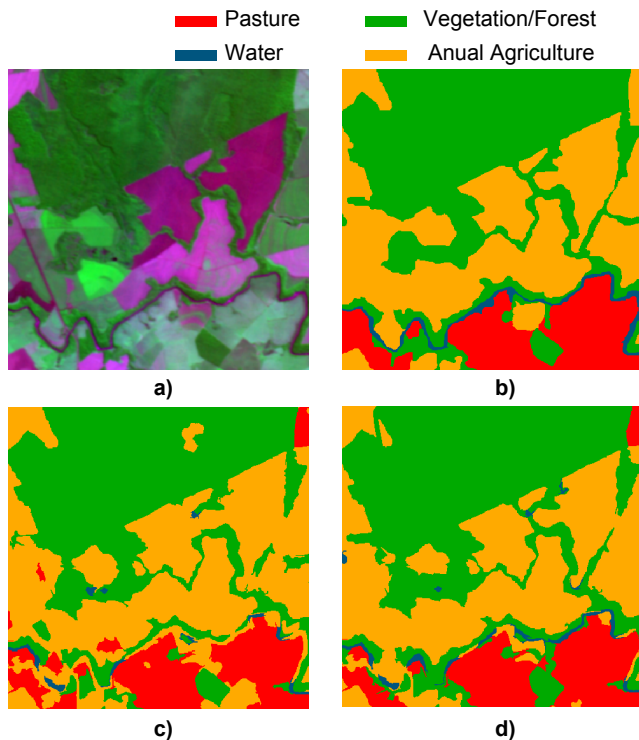


Fig. 3. Subimages of the classification map results in the CBERS dataset: a) Subimage of the CBERS image (composition: 3R, 4G, 2B). b) Subset of the ground truth. c) Subimage classification result using ACC. d) Subimage classification result using CCBOW.

5. CONCLUSION

A new scheme for contextual superpixel description based on the bag-of-visual words was proposed. The experiments performed over a multispectral image and a very high spatial resolution image show that the proposed contextual descriptors can improve classification performance as compared to widely used region descriptors. As future work, we plan to use contextual information for feature extraction and classification in a synergistic way.

Acknowledgment

This research was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (grant 131832/2013-1, 449638/2014-6), Fundação de Apoio à Pesquisa do Estado de São Paulo, Fundação de Apoio à Pesquisa do Estado de Minas Gerais (grant APQ-00768-14), and CAPES. We thank Mapmart for the QuickBird satellite imagery.

Descriptors	ROME			
	Kappa (κ)		Overall accuracy (%)	
	Mean	SD	Mean	SD
ACC	0.7709	0.0039	84.28	0.25
CCBOW	0.7921	0.0042	85.51	0.30

Table 4. Region descriptors classification performance for the ROME dataset using k-NN.

6. REFERENCES

- [1] 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*, Angers, France, May 2010, pp. 203–208.
- [2] M. Li, S. Zang, B. Zang, S. Li, and C. Wu, "A review of remote sensing image classification techniques: the role of spatio-contextual information," *European Journal of Remote Sensing*, vol. 47, pp. 389–411, 2014.
- [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 *ICPR*, Nov 2012, pp. 3090–3093.
- [4] Sheng Xu, Tao Fang, Deren Li, and Shiwei Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 2, pp. 366–370, April 2010.
- [5] A. De Giorgi, G. Moser, and S.B. Serpico, "Contextual remote-sensing image classification through support vector machines, markov random fields and graph cuts," in *IEEE International Geoscience and Remote Sensing Symposium*, July 2014, pp. 3722–3725.
- [6] Brian Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *IEEE 12th International Conference on Computer Vision*, Sept 2009, pp. 670–677.
- [7] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, Nov 2012.