CONTEXTUAL SUPERPIXEL DESCRIPTION
FOR REMOTE SENSING IMAGE CLASSIFICATION

J. E. Vargas\textsuperscript{1}, A. X. Falcão\textsuperscript{1}, J. A. dos Santos\textsuperscript{2}, J. C. D. M. Esquerdo\textsuperscript{3}, A. C. Coutinho\textsuperscript{3}, J. F. G. Antunes\textsuperscript{3}

\textsuperscript{1}Institute of Computing - University of Campinas, Campinas, Brazil
\textsuperscript{2}Department of Computer Science - Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
\textsuperscript{3}Embrapa 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 \cite{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 \cite{2}. We extend this result for superpixels using BoW — a higher-level data representation which has caught increasing attention in remote sensing image classification \cite{3, 4}. Contextual information is widely used in the classification phase \cite{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 \cite{6}, that is based on BoW and SIFT. We used Simple Linear Iterative Clustering (SLIC) algorithm \cite{7} to generate superpixels from images. In \cite{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.
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 created 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 si is adjacent to a superpixel sj if and only if a pixel from si is 4-connected to a pixel from sj. We define a contextual t-neighborhood descriptor \( D_{i}^{t} = \sum_{P(s_i, s_j) = t} D_{j}^{0} \) for superpixel si as the histogram created by aggregating histograms \( D_{j}^{0} \) from superpixels sj, where P(s_i, s_j) is the shortest-path length between si and sj 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 si 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:

\[
D_{i} = (D_{i}^{0}, D_{i}^{1}, \sum_{r \in \{2,3\}} D_{i}^{r}),
\]

where \( \sum_{r \in \{2,3\}} D_{i}^{r} \) aggregate the 2- and 3-neighborhood descriptors into one component of \( D_{i} \).

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 \( D_{i}^{t} \) 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 si in order to define its descriptor \( D_{i} \). For t = 3, for instance,

\[
D_{i} = \sum_{j \leq 3} D_{i}^{j}.
\]

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.
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.
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.

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

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Kappa (κ)</th>
<th>Overall accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>ACC</td>
<td>0.7709</td>
<td>0.0039</td>
</tr>
<tr>
<td>CCBOW</td>
<td>0.7921</td>
<td>0.0042</td>
</tr>
</tbody>
</table>

6. REFERENCES


