Contextual Description of Superpixels for Aerial Urban Scenes Classification

Tiago M. H. C. Santana*, Alexei M. C. Machado† and Jefersson A. dos Santos*
*Computer Science Department, UFMG
Belo Horizonte, Brazil
†Electrical Engineering Department, PUC Minas
Belo Horizonte, Brazil

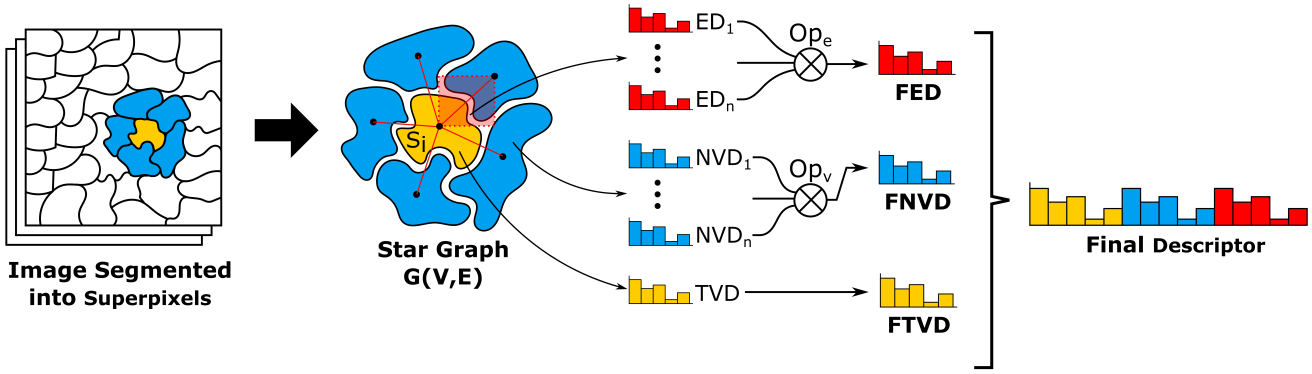


Fig. 1. Process to generate the proposed contextual descriptor for a superpixel s_i. Given a segmented image, the local neighborhood of s_i is modeled as a Region Adjacency Graph (RAG) G(V,E) in star topology where s_i is the central vertex (or the root), the superpixels adjacent to it are the leaves and edges link the mass centers of them. A feature descriptor is extracted from s_i and from each of its n neighbors. Every edge is then taken as the diagonal of a rectangle (reddish region) from which a texture descriptor is computed. The n resultant edge descriptors are combined into one of the same dimensionality through some operation Op_e. Likewise, the n neighbor vertex descriptors are used to build only one through Op_v. Lastly, the final contextual descriptor for s_i is composed by concatenating its own vertex descriptor, the final neighborhood vertex descriptor and the final edge descriptor, in this order and after individually normalizing each of them.

***Abstract*—Remote Sensing Images are one of the main sources of information about the earth surface. They are widely used to generate thematic maps that show the land cover. This process is traditionally done by using supervised classifiers which learn patterns extracted from few image pixels annotated by the user and then assign a label to the remaining pixels. However, due to the increasing spatial resolution of the images, pixelwise classification is not suitable anymore, even when combined with context. Moreover, traditional techniques used to aggregate context are unsuitable in the scenario of thematic maps generation since they depend on a previous labeling of image pixels/segments and, thus, are computationally inefficient and require a large amount of training data. Therefore, the objective of this work is to develop a description for superpixels which is able to encode their visual cues and local context without labeling them in order to generate more accurate land cover thematic maps.**

***Keywords*-contextual descriptor; land cover; thematic maps; remote sensing.**

I. Introduction

Since the Remote Sensing Images (RSIs) became available to the non-academic community, classification has played an essential role to generate new geographic products like thematic maps [1], which in turn, are fundamental for the decision-making process in several areas such as urban planning, environmental monitoring and economic activities. In this process, low-level descriptors are extracted from few image samples, such as pixels, regions, superpixels (a superpixel can be considered as a perceptually meaningful atomic region [2]), etc., which are annotated by the user, and used to train a classifier. Thereafter the generated classifier should be able to annotate the remaining samples in the image. The precision of the resultant thematic map depends on the quality of the descriptors and the training samples selected [3].

From the very beginning, RSI classification was based on pixel statistics analysis. With the increasing in the spatial resolution of the images, the information from neighboring pixels (either texture or context) was used to improve results. Although this approach has been a dominant paradigm in remote sensing for many years, pixelwise classification does not meet the current increasing demand for faster and more accurate classification anymore [4]. Region-based classification, which aims at capturing information from pixel patterns inside each segmented region of the image, has become more suitable for nowadays' scenario. Nevertheless, the use of the contextual information among regions began to be considered only very recently in RSI processing [3].

The main motivation behind using contextual information is that traditional low-level appearance features, such as color or shape, are limited while capturing the appearance variability of real world objects represented in images. In the presence of factors that modify the acquired image of a scene, such

as noise and changes in lighting conditions, the intra-class variance is increased, leading the classifier to many errors. In these scenarios, the coherent arrangement of the elements expected to be found in real world scenes can be used to help describing objects that share similar appearance features, adjusting the confidence of the classifier predictions or correct the results [5].

Existing approaches for contextual description can be divided into three categories [5]: semantic, that is regarded as the occurrence and co-occurrence of objects in scenes; scale, related to the dimension of an object with respect to the others; and spatial, that refers to the relative localization and position of objects in a scene. In addition, context can be regarded as being either global or local. The first available methods were based on fixed and predefined rules [6], [7], [8]. More effective approaches used machine learning techniques to encompass contextual relationships [9], [10], [11]. A recent trend consists on combining different kinds of context to improve the classification [12], which is nevertheless computationally inefficient and, therefore, little used so far. The main drawback of these methods is the requirement of previous identification of other elements in the image, which is computationally inefficient and requires a large amount of labeled data, making them unsuitable for automatic generation of thematic maps. A way to overcome this deficiency is through feature engineering, which consists in building a representation for image objects/regions that implicitly encodes their context instead of aggregating it as a post-processing step. This approach must somehow include co-occurrences, scale or spatial relationships between image elements without labeling them. An example can be found in the work of Lim *et al.* [13] that represents the scene as a tree of regions where the leaves are described by a combination of features from their ancestors. This resulting descriptor encodes context in a top-down fashion. To the best of our knowledge, the only approach of this type in remote sensing was proposed by Vargas *et al.* [3] to create thematic maps. In that work, each superpixel of the image is described through a histogram of visual elements, using the method of Bag of Visual Words (BoVW). Then, contextual information is encoded by concatenating the superpixel description with a combination of the histograms of its neighbors to generate a new contextual descriptor. One of the main drawbacks of this method is the lack of explicit encoding of the relational aspects among the features extracted from adjacent superpixels.

Thereby, the objective of this work is to develop a description for superpixels which is able to encode not only their visual cues, but also semantic and spatial context within a local neighborhood in order to automatically generate more accurate land cover thematic maps. Our contribution so far is a novel contextual descriptor which includes a description of the borders between superpixels. The proposed descriptor is more robust to changes in the size of the superpixels and achieves more accurate results in urban maps compared to other methods in the literature. All research done so far is reported in the following.

II. Star Descriptor

Unlike the most methods found in the literature, our approach builds a representation for image segments that implicitly encodes co-occurrences (semantic context) and spatial relations (spatial context) without the need of labeling them. The pipeline to generate the Star descriptor is summarized in Figure 1. Each step of the proposed approach is further explained in the following.

A. Segmentation Into Superpixels

Firstly, segmentation is applied to delineate objects or object parts in the image from which visual feature descriptors will be extracted. Superpixels are used instead of the traditional regions because some low level descriptors are more discriminative when extracted from regular regions such that provided by superpixel generation methods [2].

Among several methods, Simple Linear Iterative Clustering (SLIC) was chosen for this work because of it found to be more effective according to boundary recall [14]. Since the edge descriptors capture borders between adjacent superpixels, the ability of SLIC to adhere to the borders of objects in the image can leverage edge descriptors computed by our descriptor.

B. Graph Modeling

Given an image segmented into N superpixels, the local neighborhood of each superpixel s_i, $i = 1, \ldots, N$, is regarded as a Region Adjacency Graph (RAG) $G(V, E)$ in star topology (see Figure1) where V are the superpixels and the edges in E represent adjacency relation between s_i and the other superpixels. Formally, two superpixels s_x and s_y are adjacent if and only if at least one pixel of s_x is 4-connected to a pixel of s_y. In addition, the target superpixel s_i is the central vertex (or root), each of its n neighbor superpixels ns_j, $j = 1, \ldots, n$, are the leaves and there is an edge e_k, $k = 1, \ldots, n$ linking the mass centers of s_i and every ns_j. Such a graph modeling provides a clear understanding of the proposed descriptor in terms of the level of context taken into account and the types of context exploited (spatial relations and co-occurrences between the s_i and a pattern of neighborhood).

C. Vertex Descriptors

A visual feature descriptor is computed within every superpixel in a given local neighborhood modeled as a RAG $G(V, E)$. More formally, a feature vector - referred to as target vertex descriptor (TVD) - is extracted from the target superpixel s_i. Likewise, a neighbor vertex descriptor (NVD_j) is built for ns_j, $j = 1, \ldots, n$, as can be seen in Figure 1. Notice that the same algorithm is used for both TVD and every NVD_j.

Although the only restriction for the vertex descriptor chosen is that it must represent every superpixel by a fixed-size numerical vector, we propose to use two types: low level global color/texture descriptors and BoVW for mid level representation. In the former approach, a global descriptor is extracted from each superpixel taking it as it were a

whole image. To account for size differences among them, the resultant feature vector is normalized. The second way is explained in the work of Vargas *et al.* [3].

D. Edge Descriptors

The edge descriptor proposed by Silva *et al.* [15] was used to better capture the patterns found in the borders of neighbors, since it directly represents the transition across the frontiers of two adjacent superpixels by extracting texture descriptors around the edge. More precisely, given a local neighborhood represented as a RAG, the *k*-th edge descriptor (ED_k) is computed by extracting a low level texture descriptor within the rectangle formed by taking e_k as its diagonal (as exemplified by the reddish area nearby the edge in Figure 1). This process is repeated for each of the n edges in E.

E. Final Descriptor Composition

Since the vertex and edge descriptors were extracted, they are combined into only one vertex descriptor and one edge descriptor through some operation. This step is applied to tackle with two issues: due to the large number of feature vectors extracted from each RAG, the computational cost to train a classifier with them would be prohibitively high and the variability in the number of leaves of the graphs would result in a feature vector of non-fixed size if a simple concatenation would be done.

More specifically, an operation Op_v is applied to summarize the n NVDs, resulting in one final neighbor vertex descriptor ($FNVD$). Similarly, the n EDs are combined into just one final edge descriptor (FED) through an operation Op_e. The final target vertex descriptor ($FTVD$) is the TVD itself. Because vertex and edge descriptors lie in different feature spaces, $FTVD$, $FNVD$ and FED are individually normalized using L_2 norm and then concatenated to compose the final descriptor which has $2 * |vertexdescriptor| + |edgedescriptor|$ dimensions.

The only constraint imposed to Op_v and Op_e is that they must summarize p *m*-dimensional vectors into one of same dimensionality. Concretely, we propose to use three operations commonly found in BoVW pooling step: sum pooling, average pooling and max pooling. These operations are formally defined as follows: let D_j be the *j*-th *m*-dimensional feature vector in a sequence $\langle D_1, \ldots, D_p \rangle$, whose components are d_i, $i \in \{1, \ldots, m\}$ as stated in Equation 1; the *i*-th component of D_j can be summarized through either sum, average or max pooling, which are respectively showed in Equation 2.

$$D_j = \{d_i\}_{i \in \{1,\ldots,m\}} \tag{1}$$

$$d_i = \sum_{j=1}^{p} d_{i,j} \qquad d_i = \frac{1}{p} \sum_{j=1}^{p} d_{i,j} \qquad d_i = \max_{j \in \{1,\ldots,p\}} d_{i,j} \tag{2}$$

III. Experiments and Results

Datasets. The experiments were carried out on two imbalanced multi-class datasets: the grss_dfc_2014 and ISPRS. The first dataset consists of a Very High Resolution (VHR) image, spatial resolution of 20 cm, taken over an urban area near Thetford Mines in Québec, Canada. The ground truth was annotated into seven classes. The grss_dfc_2014 dataset provides a specific subset of the entire image for training a classifier which should be used to generate a thematic map for the whole image. The ISPRS consists of 38 very high resolution true orthophoto (TOP) image patches of 5 cm of ground sampling distance, taken over Postdam, Germany. The ground truth is provided for 24 image patches annotated into 6 classes. Due to the large amount of descriptors extracted by Star descriptor, we randomly selected 5 in the 24 annotated images to perform the experiments: 3_12, 4_12, 5_12, 7_11 and 7_12. **Setup.** The first experiment aims at identifying the best configuration of the Star descriptor on the grss_dfc_2014. The superpixel segmentation was performed using SLIC with 25,000 regions and 25 of compactness for the training image and 37,000 regions and 25 of compactness for the whole image. The number of regions for the whole image was chosen to be 37,000 because the image is about 50% bigger than the training image. The vertices were described by using one texture - Unser (USR) - and three color descriptors - Border/Interior pixel Classification (BIC), Color Coherence Vector (CCV) and Global Color Histogram (GCH) - as either global descriptor or BoVW with 256 words in the codebook. Histograms of Local Binary Patterns (LBP) and USR descriptor were computed for the edges. All three operations - sum, average and max pooling - were used to summarize the final vertices and edge descriptor. The extracted contextual descriptors were used to train a Support Vector Machine (SVM) classifier with Radial Basis Function (RBF) kernel and the parameters were determined through grid searching 5-fold Cross-validation in the training set. In order to assess the robustness of Star descriptor to changes in the segmentation scale, a second experiment was carried out varying the number of regions of SLIC for the best configuration of the first experiment. The third experiment is similar to the first one, but carried out on the ISPRS dataset. The 5 randomly selected images were used to perform a 5-fold Cross-validation. Each image was segmented using SLIC with 32000 regions and 25 of compactness. BIC was used as global descriptor to describe vertices and LBP for the edges. Sum pooling was used with the baseline and max pooling with Star, since these are the original proposals of the methods. The classifier and parameter search remains the same. **Baselines.** The first baseline is the low/mid level representation for the superpixels without any context. The second baseline is the contextual descriptor proposed by Vargas *et al.* [3] which is the only approach that implicitly encodes context with the purpose of generating thematic maps. **Evaluation metrics.** All results are reported in terms of overall accuracy (Ovr.) and Kappa index (κ). The last experiment is evaluated using Student's *t* distribution with 97.5% of confidence. It is worth to mention that although a single label is assigned to each superpixel, the metrics are calculated in terms of pixels. This is done by assigning the label of the superpixel to every pixel within it. **Results.** After carrying out the first experiment, the configuration of the proposed descriptor which achieved the best results includes BIC without BoVW as vertex descriptor,

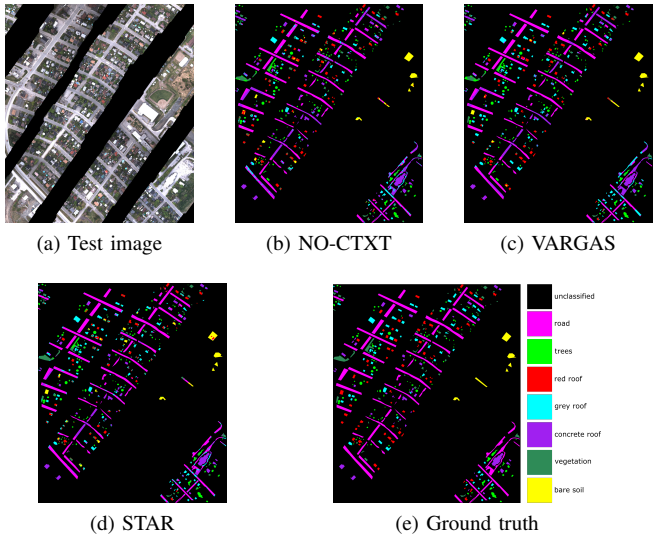
TABLE I
COMPARISON BETWEEN STAR DESCRIPTOR AND BASELINES

	grss_dfc_2014		ISPRS	
Descriptors	κ	Ovr.	κ	Ovr.
NO-CTXT	0.619	0.724	0.230±0.040	0.474±0.065
VARGAS	0.651	0.751	**0.275±0.053**	**0.501±0.058**
STAR	**0.735**	**0.822**	0.181±0.049	0.421±0.064

USR as edge descriptor and max and average pooling to summarize them, respectively. The baseline without context using BIC with BoVW and the proposal of Vargas *et al.* using BIC as global descriptor and max pooling found to be the best configurations. A comparison among them is reported in Table I, under the names STAR, NO-CTXT and VARGAS, respectively. The resultant maps are shown in Figure 2.

Results of the second experiment are presented in Figure 3. As can be seen from the graphic, Star descriptor is more robust to changes in segmentation scale than Vargas' descriptor, whose Kappa index drastically drops for more than 36,997 regions, becoming worse than the baseline without context.

Results of the third experiment are also reported in Table I. Notice that Star descriptor was the worse in these preliminary results because we have not found its best configuration for the ISPRS dataset yet.



(a) Test image (b) NO-CTXT (c) VARGAS

(d) STAR (e) Ground truth

Fig. 2. Thematic maps generated

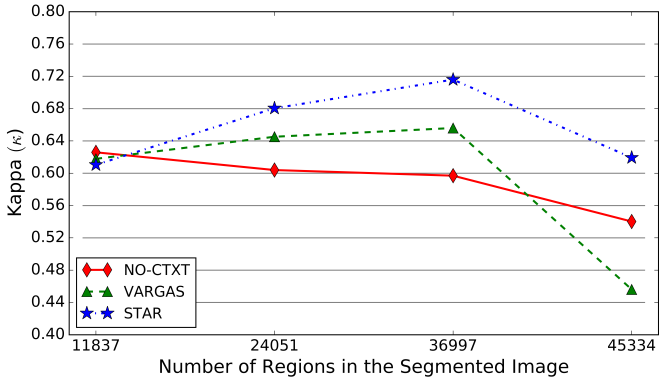


Fig. 3. Comparison of robustness of descriptors to changes in the segmentation scale

IV. CONCLUSION AND NEXT STEPS

A new approach for superpixel description which encodes context was presented in this work. Unlike most methods in the literature, Star descriptor does not require strongly labeled data to aggregate context, which makes it suitable for thematic maps creation. Our approach was more robust to changes in the size of superpixels and created more accurate thematic maps in the experiments carried out on grss_dfc_2014 dataset. These results show that encoding context improves classification results and borders might be important in urban scenes. Experiments to find the best configuration of Star in the ISPRS dataset are still in progress. Besides that, the next steps include the usage of a pre-trained convolutional network to extract features of the borders between superpixels.

ACKNOWLEDGEMENTS

This work was partially financed by CNPq, CAPES, Fapemig. The authors would like to thank Telops Inc. (Québec, Canada) for acquiring and providing the data used in this study, the IEEE GRSS Image Analysis and Data Fusion Technical Committee and Dr. Michal Shimoni (Signal and Image Centre, Royal Military Academy, Belgium) for organizing the 2014 Data Fusion Contest, the Centre de Recherche Public Gabriel Lippmann (CRPGL, Luxembourg) and Dr. Martin Schlerf (CRPGL) for their contribution of the Hyper-Cam LWIR sensor, and Dr. Michaela De Martino (University of Genoa, Italy) for her contribution to data preparation.

REFERENCES

[1] G. G. Wilkinson, "Results and implications of a study of fifteen years of satellite image classification experiments," *IEEE T. Geosci. Remote*, vol. 43, no. 3, pp. 433–440, 2005.

[2] R. Achanta, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," EPFL, Tech. Rep., June 2010, technical Report 149300.

[3] J. E. Vargas, A. X. Falcão, J. A. dos Santos, J. C. D. M. Esquerdo, C. A. C., and J. F. G. Antunes, "Contextual superpixel description for remote sensing image classification," in *Proceedings of Int. Geosci. Remote Se.* IEEE, 2015.

[4] T. Blaschke, G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Q. Feitosa, F. van der Meer, H. van der Werff, F. van Coillie, and D. Tiede, "Geographic object-based image analysis towards a new paradigm," *ISPRS J. Photogramm.*, vol. 87, pp. 180 – 191, 2014.

[5] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Comput. Vis. Image Und.*, vol. 114, no. 6, pp. 712–722, Jun. 2010.

[6] A. R. Hanson and E. M. Riseman, "VISIONS: A computer system for interpreting scenes," in *Computer Vision Systems*, A. R. Hanson and E. M. Riseman, Eds. New York: Academic Press, 1978.

[7] T. M. Strat and M. A. Fischler, "Context-based vision: recognizing objects using information from both 2d and 3d imagery," *IEEE T. Pattern Anal.*, vol. 13, no. 10, pp. 1050–1065, Oct 1991.

[8] M. A. Fischler and R. Elschlager, "The representation and matching of pictorial structures," *IEEE T. Comput.*, vol. C-22, no. 1, pp. 67–92, Jan 1973.

[9] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *IEEE I. Conf. Comp. Vis.*, Oct 2007, pp. 1–8.

[10] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[11] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vision*, vol. 81, no. 1, pp. 2–23, Jan. 2009.

[12] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proc. CVPR IEEE*, June 2014, pp. 891–898.

[13] J. J. Lim, P. Arbelaez, P. Arbelez, C. Gu, and J. Malik, "Context by region ancestry," in *IEEE I. Conf. Comp. Vis.*, Sept 2009, pp. 1978–1985.

[14] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE T. Pattern Anal.*, vol. 34, no. 11, pp. 2274–2282, Nov 2012.

[15] F. B. Silva, S. Goldenstein, S. Tabbone, and R. d. S. Torres, "Image classification based on bag of visual graphs," in *IEEE Image Proc.*, Sept 2013, pp. 4312–4316.