

Star: A Contextual Description of Superpixels for Remote Sensing Image Classification

Tiago M.H.C. Santana¹(✉), Alexei M.C. Machado², Arnaldo de A. Araújo¹,
and Jefersson A. dos Santos¹

¹ Department of Computer Science, Universidade Federal de Minas Gerais (UFMG),
Belo Horizonte, Brazil
{tiago@dcc.ufmg.br,arnaldo@dcc.ufmg.br,jefersson@dcc.ufmg.br}

² Department of Computer Science, Pontifícia Universidade Católica de Minas Gerais,
Belo Horizonte, Brazil
alexei@pucminas.br

Abstract. Remote Sensing Images are one of the main sources of information about the earth surface. They are widely used to automatically generate thematic maps that show the land cover of an area. This process is traditionally done by using supervised classifiers which learn patterns extracted from the 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 resulting from advances in the acquisition technology, pixelwise classification is not suitable anymore, even when combined with context. Therefore, we propose a new descriptor for superpixels called Star descriptor that creates a representation based on both its own visual cues and context. Unlike the most methods in the literature, the new approach does not require any prior classification to aggregate context. Experiments carried out on urban images showed the effectiveness of the Star descriptor to generate land cover thematic maps.

Keywords: Remote sensing · Thematic maps · Land cover · Contextual descriptor

1 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, texture 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 [6–8]. More effective approaches used machine learning techniques to encompass contextual relationships [9–11]. 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. 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. This approach must somehow include co-occurrences, scale or spatial relationships between descriptors of 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, this paper presents a novel contextual descriptor for superpixels of RSIs, which builds a representation for the target superpixel in terms of its own visual cues, visual features extracted from its neighbors and pairwise interactions between them. The resulting representation implicitly encodes spatial relationships and co-occurrences of patterns extracted within a neighborhood and, thus, does not require any prior classification to aggregate context.

2 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 Fig. 1. Each step of the proposed approach is further explained in the following.

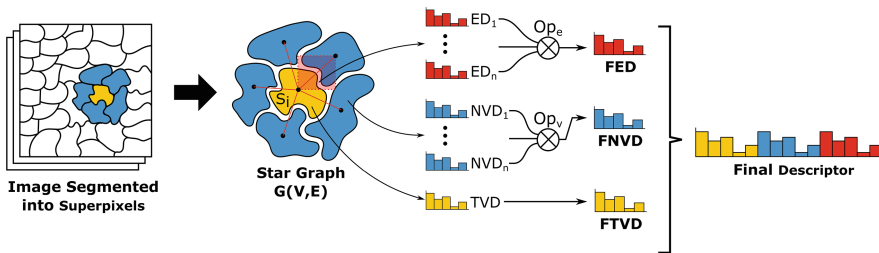


Fig. 1. Process to generate the proposed contextual descriptor for a superpixel s_i . Given a segmented image, the local neighborhood of s_i is modeled as a star graph $G(V,E)$ where s_i is the central vertex (or the root), the superpixels adjacent to it are the leaves and edges link the mass centers 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. (Color figure online)

2.1 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.

2.2 Graph Modeling

Given an image segmented into N superpixels, the local neighborhood of each superpixel s_i , $i = 1, \dots, N$, is regarded as a graph $G(V, E)$ in star topology (see Fig. 1) 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 , $j = 1, \dots, n$, are the leaves and there is an edge e_k , $k = 1, \dots, n$ 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).

2.3 Vertex Descriptors

A visual feature descriptor is computed within every superpixel in a given local neighborhood modeled as a star graph $G(V, E)$. 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 , $j = 1, \dots, n$, as can be seen in Fig. 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 was proposed by Vargas *et al.* [3]: dense grid sampling is applied and low level color/texture descriptors are computed from each 5×5 local patch around the selected pixels; the extracted feature vectors are used to conform the codebook using the k -means clustering algorithm; hard assignment is used to assign the closest visual word to each pixel of the grid; a histogram is then computed for every superpixel by taking into account the grid pixels within it; finally, a normalization is applied to each histogram, which is divided by the number of grid pixels inside its respective superpixel.

2.4 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 star graph $G(V, E)$, 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 Fig. 1). This process is repeated for each of the n edges in E .

2.5 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 star graph, 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 $\langle D_1, \dots, D_p \rangle$, whose components are d_i , $i \in \{1, \dots, m\}$ as stated in Eq. 1; the i -th component of D_j can be summarized through either sum, average or max pooling, which are respectively showed in Eq. 2.

$$D_j = \{d_i\}_{i \in \{1, \dots, m\}} \quad (1)$$

$$d_i = \sum_{j=1}^p d_{i,j} \quad d_i = \frac{1}{p} \sum_{j=1}^p d_{i,j} \quad d_i = \max_{j \in \{1, \dots, p\}} d_{i,j} \quad (2)$$

3 Experimental Protocol

Datasets. The experiments were carried out on an imbalanced multi-class dataset: the *grss_dfc_2014* [16]. The dataset consists of a Very High Resolution (VHR) image, spatial resolution of 20 cm, taken in 2013 over an urban area near Thetford Mines in Québec, Canada. This dataset was annotated into seven classes: road, trees, red roof, grey roof, concrete roof, vegetation and bare soil. 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. **Setup.** 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. Since color descriptors usually achieve better results in RSI classification [17], the vertices were described by using just one texture - Unser (USR) [17]- and three color descriptors - Border/Interior pixel Classification (BIC), Color Coherence Vector (CCV) and Global Color Histogram (GCH) [17] - as either global descriptor or BoVW with 256 words in the codebook. Histograms of Local Binary Patterns (LBP) were used initially for the edges because it is the original proposal of Silva *et al.* [15]. However, LBP is not one of the best options for RSI classification [17]. For this reason, USR (which is a good trade-off between accuracy and number of dimensions) was also tested. 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 configurations of the first experiment. **Baselines.** The first baseline used for comparison is the low/mid level representation for the superpixels without any context, which is referred to as NO-CTXT in Sect. 4. 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. Its results are shown under the name VARGAS in the next section. It is worth to mention that the results reported for NO-CTXT used BIC with BoVW and the configuration for VARGAS consists of BIC as global descriptor, which achieved the best accuracy for each of them. **Evaluation metrics.** All results are reported in Sect. 4 in terms of overall accuracy (Ovr.), average accuracy (Avg.) and Kappa index (κ). 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.

4 Results

The best results achieved by Star descriptor for each vertex descriptor are reported in Table 1 for LBP and USR as edge descriptors. As can be seen from the table, BIC was the most prominent vertex descriptor when combined with max pooling. In general, the average pooling operation was better to summarize the edge descriptor and the USR descriptor obtained the best results for the proposed descriptor, shown in boldface in Table 1. Notice that although using BoVW usually produces slightly more accurate maps, the highest kappa index was achieved using BIC as a global descriptor.

A comparison between the best configuration of Star found in Table 1 and the baselines is presented in Table 2. The proposed contextual descriptor achieved

Table 1. Best results of Star descriptor on grss_dfc_2014

Edge descriptor:		LBP					USR				
Global/BoVW	Descriptors	Op_v	Op_e	κ	Ovr.	Avg.	Op_v	Op_e	κ	Ovr.	Avg.
BoVW	STAR-BIC	Max	Avg	0.681	0.772	0.812	Max	Avg	0.707	0.795	0.805
	STAR-CCV	Max	Avg	0.638	0.740	0.769	Max	Max	0.679	0.779	0.756
	STAR-GCH	Max	Sum	0.642	0.746	0.757	Max	Avg	0.680	0.780	0.753
	STAR-USR	Max	Avg	0.559	0.684	0.648	Max	Avg	0.522	0.655	0.623
Global	STAR-BIC	Max	Avg	0.522	0.664	0.592	Max	Avg	0.735	0.822	0.775
	STAR-CCV	Max	Avg	0.636	0.742	0.753	Avg	Sum	0.676	0.782	0.737
	STAR-GCH	Max	Avg	0.630	0.738	0.743	Max	Avg	0.677	0.779	0.745
	STAR-USR	Sum	Max	0.555	0.684	0.629	Sum	Sum	0.510	0.648	0.595

better results than all baselines for all metrics used to assess them. Another key observation is that encoding context to describe superpixels improved the accuracy of the automatic generated thematic maps. It is worthwhile to mention that in the experiments carried out, Star descriptor usually performed better than baselines when combined with the most low/mid level descriptors and operations that summarize them.

Results of the second experiment are presented in Fig. 2. 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.

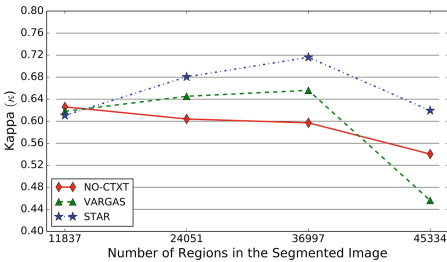


Fig. 2. Comparison of robustness of descriptors to changes in the segmentation scale

Table 2. Comparison between STAR descriptor and baselines

Descriptors	κ	Ovr.	Avg.
NO-CTXT	0.619	0.724	0.767
VARGAS	0.651	0.751	0.766
STAR	0.735	0.822	0.775

5 Conclusion

A new approach for superpixel description which encodes context was proposed in this paper. The Star descriptor builds a representation for each superpixel in terms of its own visual cues and in terms of its context, by taking into account

the spatial relations and co-occurrences of visual patterns within the local neighborhood, modeled as a star graph. A classifier is then trained with the resultant feature vectors of the training set and used to generate a thematic map by painting the remaining superpixels according to the labels assigned to them. From the experiments carried out on the `grss_dfc_2014` dataset, we found that the use of context improved the quality of the resultant maps and the proposed descriptor achieved better results over the baselines. We believe that it is possible to generate even more representative contextual descriptors by using combinations of low-level descriptors or other operations which preserves the information about the relative positions of the superpixels. In future, we plan to investigate these improvements and assess the Star descriptor using learned features as vertex descriptors.

Acknowledgements. This work was partially financed by CNPq, CAPES, and 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. Wilkinson, G.G.: Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Trans. Geosci. Remote* **43**(3), 433–440 (2005)
2. Achanta, R., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels. Technical report 149300, EPFL, June 2010
3. Vargas, J.E., Falcão, A.X., dos Santos, J.A., Esquerdo, J.C.D.M., Coutinho, A.C., Antunes, J.F.G.: Contextual superpixel description for remote sensing image classification. In: *International Geoscience and Remote Sensing Symposium*. IEEE (2015)
4. Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D.: Geographic object-based image analysis towards a new paradigm. *ISPRS J. Photogramm.* **87**, 180–191 (2014)
5. Galleguillos, C., Belongie, S.: Context based object categorization: a critical survey. *Comput. Vis. Image Underst.* **114**(6), 712–722 (2010)
6. Hanson, A.R., Riseman, E.M.: VISIONS: a computer system for interpreting scenes. In: Hanson, A.R., Riseman, E.M. (eds.) *Computer Vision Systems*. Academic Press, New York (1978)
7. Strat, T.M., Fischler, M.A.: Context-based vision: recognizing objects using information from both 2D and 3D imagery. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(10), 1050–1065 (1991)
8. Fischler, M.A., Elschlager, R.: The representation and matching of pictorial structures. *IEEE Trans. Comput.* **C-22**(1), 67–92 (1973)

9. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: IEEE 11th International Conference on Computer Vision, pp. 1–8, October 2007
10. Torralba, A.: Contextual priming for object detection. *Int. J. Comput. Vis.* **53**(2), 169–191 (2003)
11. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. J. Comput. Vis.* **81**(1), 2–23 (2009)
12. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the CVPR IEEE, pp. 891–898, June 2014
13. Lim, J.J., Arbelaez, P., Arbelaz, P., Gu, C., Malik, J.: Context by region ancestry. In: IEEE International Conference on Computer Vision, pp. 1978–1985, September 2009
14. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(11), 2274–2282 (2012)
15. Silva, F.B., Goldenstein, S., Tabbone, S., Torres, R. da S.: Image classification based on bag of visual graphs. In: IEEE International Conference on Image Processing, pp. 4312–4316, September 2013
16. IEEE: GRSS data fusion contest (2014). <http://www.grssiieee.org/community/technical-committees/data-fusion/>
17. dos Santos, J.A., Penatti, O.A.B., Torres, R. da S.: Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification. In: VISAPP, Angers, France, May 2010