

SCENE CATEGORIZATION USING BAG OF TEXTONS ON SPATIAL HIERARCHY

S. Battiato, G. M. Farinella, G. Gallo, D. Ravi

{battiato, gfarinella, gallo, ravi}@dmi.unict.it

Dipartimento di Matematica e Informatica
Università di Catania
Viale Andrea Doria 6, 95125, Catania

ABSTRACT

This paper proposes a method to recognize scene categories using bags of visual words obtained hierarchically partitioning into subregion the input images. Specifically, for each subregions the Texton histogram and the extension of the subregion is taken into account. The bags of visual words, obtained in this way, are weighted and used in a similarity measure during the categorization. Experimental tests using ten different scene categories show that the proposed approach achieves good performances with respect to the state of the art methods.

Index Terms— Scene Categorization, Bag of Visual Words, Textons, Spatial Representation

1. INTRODUCTION

Automatic scene categorization is useful in many relevant computer vision applications such as content-based image retrieval (CBIR) [1] or bootstrap learning to select the advertising to be sent by Multimedia Messaging Service (MMS) [2].

Existing methods works extracting local concepts directly on spatial domain [1, 3] or in frequency domain [4]. A global representation of the scene is obtained grouping together local information in different ways (e.g., histogram of visual concepts, spectra template, etc.). Recently, the spatial layout of the local information have been used to improve the classification quality [5]. Typically, memory-based recognition algorithms (e.g., Support Vector Machine, K-Nearest Neighbor, etc.) are employed, together with holistic representation of the scene, to assign the scene category skipping the recognition of the objects that are present in the scene [4].

In this paper we propose to recognize scene categories by means of bags of visual words [6]. These are computed after hierarchically partitioning the images in subregions. Specifically, each subregion is represented as a histograms of Textons [7]. To every histogram a weight, inversely proportional to the extension of the related subregion, is assigned and it is used in the computation of similarity. Like in [5] we penalize

histograms related to larger regions because they can involve increasingly dissimilar visual words.

The proposed approach has been experimentally tested on a large database of about 4000 images. Ten different basic categories of scene have been considered. In spite of the simplicity of the proposal, the results are promising: the categorization accuracy obtained, closely matches the results of other state-of-the-art solutions [3, 4, 5].

The rest of the paper is organized as follows: Section 2 describes the model we have used for representing the images. Section 3 illustrates the dataset, the setup involved in our experiments and the results obtained using the proposed approach. Finally, in Section 4 we conclude with avenues for further research.

2. WEIGHTING BAGS OF TEXTONS

Scene categorization is typically performed describing images through feature vectors encoding color, texture, and other visual properties such as corners, edges, local interest points, etc. These information can be automatically extracted using several image processing algorithms and represented by many different local descriptors. A holistic global representation of the scene is built grouping together such local information. This representation is then used during categorization (or retrieval) tasks.

Local features denote distinctive patterns and properties of the region from which have been generated. In recent works, authors suggest to consider these patterns as “visual words” [6]: an image may hence be considered as a bag of such words.

To use the bag of “visual words” model, a visual vocabulary is built during the learning phase: all the local features extracted from the training images are clustered. The prototype of each cluster is treated as a “visual word” representing a special local pattern. This is the pattern sharing the main distinctive properties of the local features within the cluster. In this manner a visual-word vocabulary is built.

Through this process, all images from the training and the

test sets may be considered as “document” composed of “visual words” from a finite vocabulary. Indeed, each local feature within an image is associated to the closest visual word within the built vocabulary. This intermediate representation is then used to obtain a global descriptor. Typically, the global descriptor encodes the frequencies of each visual word within the image under consideration.

As pointed out in [5], this type of approach leaves out the information about the spatial layout of the local features. Differently than in text documents, spatial layout of local features for images is crucial. The relative position of a local descriptor can help in disambiguate concepts that are similar in terms of local descriptor. For instance, sky and sea could be similar in terms of local descriptor, but are typically different in terms of position within the scene.

To overcome these difficulties we propose to augment the basic bag of visual words representation combining it with a hierarchical partitioning of the image. Differently than in [5], where a single regular grid is employed, we partition an image using three different modalities: horizontal, vertical, regular grid. These schemes are recursively applied to obtain a hierarchy of subregions as shown in Figure 1.

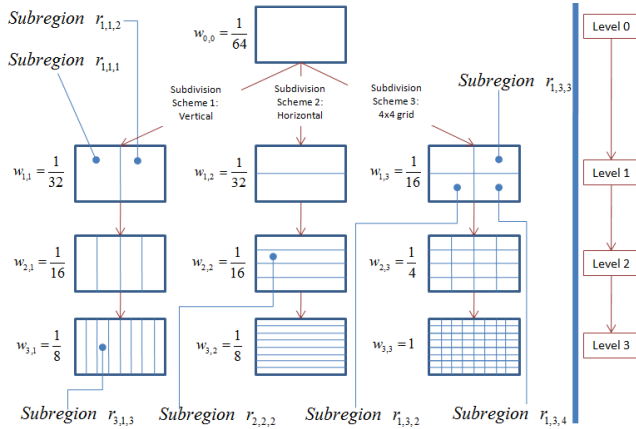


Fig. 1. Subdivision schemes up to the fourth hierarchical levels. The i_{th} subregion at level l in the subdivision scheme s is identified by $r_{l,s,i}$. The weights $w_{l,s}$ are defined by the Equation (1).

The bag of visual words representation is hence computed in the usual way, using a pre-built vocabulary, relatively to each subregion in the hierarchy. In this way we take into account the spatial layout information of local features. The proposed augmented representation hence, keeps record of the frequency of the visual words in each subregion (Figure 2).

A similarity measure between images may now be defined as follows. First, a similarity measure between histograms of visual words relative to corresponding regions is computed. The choice of such measure is discussed in Section 2.2. The connection of similarity values of each subregion are then

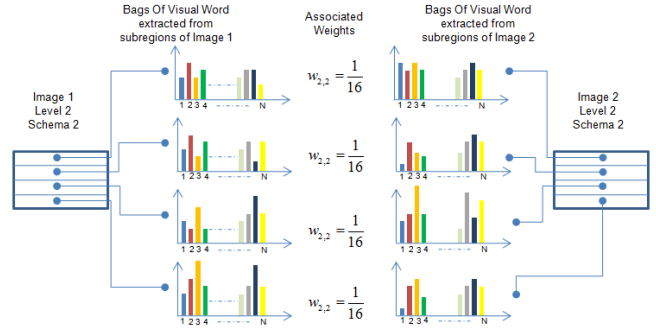


Fig. 2. A toy example of the similarity evaluation between two images I_1 and I_2 at level 2 of the subdivision schema 2. After representing each subregion $r_{2,2,i}^I$ as a distribution of Textons $B(r_{2,2,i}^I)$, the distance $D_{2,2}(I_1, I_2)$ between the two images is computed taking into account the defined weight $w_{2,2}$.

combined into a final distance by means of a weighted sum. The choice of weight is justified by the following rationale: the probability to find a specific visual word in a subregion at fine resolution is sensibly lower than finding the same visual word in a subregion with higher resolution. We penalize similarity in larger subregion defining weights inversely proportional to the subregions size (Figure 1, Figure 2).

Formally, denoting with $S_{l,s}$ the number of subregions at level l in the scheme s , the distances between corresponding subregions of two different images considered at level l in the scheme s is weighted as follows:

$$w_{l,s} = \frac{S_{l,s}}{\max_{Level, Scheme}(S_{Level, Scheme})} \quad (1)$$

where $Level$ and $Scheme$ span on all the possible level and schemas involved in a predefined hierarchy.

In the following subsections we provide more details about the local features used to build the bag of visual words representation and other aspects of the overall categorization process.

2.1. Local Feature Extraction

Previous studies emphasize the fact that global representation of scenes based on extracted holistic cues can effectively help to solve the problem of rapid and automatic scene classification [4]. Textons are distinctive image patterns [7]. We choose to use Textons as the visual words able to identify properties and structures of different textures present in the scene. To build the visual vocabulary each image in the training set is processed with a bank of filters. All responses are then clustered, pointing out the Textons vocabulary, by considering the cluster centroids. Each image pixel is then associated to the closest Texton taking into account its filter bank responses.

More precisely good results have been achieved by considering a bank of 2D Gabor filters and the k-means clustering to build the Textons vocabulary. Each pixel has been associated with a 24-dimensional feature vector obtained processing each gray scaled image through 2D Gabor filters:

$$G(x, y, f_0, \theta, \alpha, \beta) = e^{-(\alpha^2 x'^2 + \beta^2 y'^2)} \times e^{j2\pi f_0 x'} \quad (2)$$

$$x' = x \cos \theta + y \sin \theta \quad (3)$$

$$y' = -x \sin \theta + y \cos \theta \quad (4)$$

The 24 Gabor filters have size 49×49 , obtained considering two different frequencies of the sinusoid ($f_0 = 0.33, 0.1$), three different orientations of the Gaussian and sinusoid ($\theta = -60^\circ, 0, 60^\circ$), two different sharpness of the Gaussian major axis ($\alpha = 0.5, 1.5$) and two different sharpness of the Gaussian minor axis ($\beta = 0.5, 1.5$). Each filter is centered at the origin and no phase-shift is applied. We preliminarily tested other bank of filters. The best results have been obtained employing the Gabor bank of filters as discussed above.

The experiments reported in Section 3 are performed on spatial hierarchy involving the three subdivision schemes (Figure 1) with three levels ($l = 0, 1, 2$) and employing a visual vocabulary of 400 Textons.

2.2. Distance Between Images

As we pointed out previously the weighted distance that we propose to use is founded on the similarity between two corresponding subregions when the bag of visual words have been computed on the same vocabulary.

Let $B(r_{l,s,i}^{I_1})$ and $B(r_{l,s,i}^{I_2})$ the bags of visual words representation of the i_{th} subregion at level l in the schema s of two different images I_1 and I_2 . We use the metric based on Bhattacharyya Coefficient to measure the distance between $B(r_{l,s,i}^{I_1})$ and $B(r_{l,s,i}^{I_2})$. Such distance measure has several desirable properties [8]: it imposes a metric structure, it has a clear geometric interpretation, it is valid for arbitrary distributions, it approximates the χ^2 statistic avoiding the singularity problem of the χ^2 test when comparing empty histogram bins.

The distance between two images I_1 and I_2 at level l of the schema s is:

$$D_{l,s}(I_1, I_2) = w_{l,s} * \sum_i \sqrt{1 - \rho[B(r_{l,s,i}^{I_1}), B(r_{l,s,i}^{I_2})]} \quad (5)$$

$$\rho[B(r_{l,s,i}^{I_1}), B(r_{l,s,i}^{I_2})] = \sum_T \sqrt{B(r_{l,s,i}^{I_1})_T * B(r_{l,s,i}^{I_2})_T} \quad (6)$$

where $B(r_{l,s,i}^I)_T$ indicate the frequency of a specific Texton T in the subregion $r_{l,s,i}$ of the image I .

Observe that the level 0 of the hierarchy corresponds to the classic bag of visual words model in which the metric based on Bhattacharyya Coefficient is used to establish the distance between two images.

The final distance between two images I_1 and I_2 is hence computed as follows:

$$D(I_1, I_2) = D_{0,0} + \sum_l \sum_s D_{l,s} \quad (7)$$

This distance is coupled with a K-Nearest Neighbor (KNN) algorithm in order to recognize the class of a scene.

3. EXPERIMENTS AND RESULTS

The dataset we have used contains more than 4000 images collected by the authors in [3, 4, 5]. We grouped these images in 10 basic categories of scenes (Figure 3): Coast, Forest, House, Suburban, Office, Open Countries, Mountains, Buildings, Store, Highway. Moreover, these basic categories can be ensembled and described with a major level of abstraction (Figure 3): In vs. Out, Natural vs. Artificial. The categorization task is realized through a simple KNN classifier using the representation and similarity measure described in Section 2.

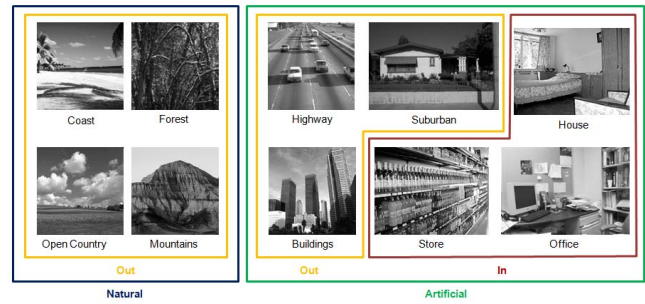


Fig. 3. Some examples of images used in our experiments considering basic and superordinate level of description.

	Suburban	Store	Buildings	Highway	Mountains	Open Country	Coast	Forest	Office	House
Suburban	98,29	0,00	0,57	0,00	0,00	0,00	0,00	0,57	0,00	0,57
Store	9,35	75,83	2,16	0,00	0,00	0,00	0,00	7,48	0,43	4,75
Building	4,48	13,70	62,34	2,33	1,33	3,48	0,54	5,38	0,65	5,77
Highway	2,28	0,00	1,14	88,03	0,00	3,99	2,85	1,14	0,00	0,57
Mountains	2,40	1,13	2,41	3,81	45,06	31,36	5,08	8,33	0,00	0,42
Open Country	0,58	0,00	1,01	4,60	1,58	75,68	11,51	5,04	0,00	0,00
Coast	0,43	0,00	0,85	9,53	1,70	27,69	58,95	0,85	0,00	0,00
Forest	0,43	0,86	0,00	0,00	0,86	4,71	0,00	93,14	0,00	0,00
Office	5,25	1,42	1,42	0,00	0,00	0,00	0,00	0,00	76,03	15,88
House	4,09	9,13	3,09	1,14	0,14	1,20	0,00	1,00	8,84	71,37

Table 1. Confusion Matrix obtained considering the proposed approach on the basic level of description of the scenes. The average classification rates for individual classes are listed along the diagonal.

	Natural	Artificial
Natural	92,88	7,12
Artificial	5,98	94,02

Table 2. Natural vs. Artificial categorization.

	Out	In
Out	91,63	8,37
In	11,50	88,50

Table 3. In vs. Out categorization

All categorization experiments have been repeated ten times by using a ten-fold cross validation approach. The per-class classification rates were recorded at each run in a confusion matrix in order to evaluate the performances of the model at each run. The averages from the individual runs are reported through confusion matrices in Tables 1, 2, 3 (the x-axis represents the inferred classes while the y-axis represents the ground-truth category). The overall classification rate is 75% considering the ten basic classes, 90,06% considering the superordinate level of description In vs. Out, 93,4% considering the superordinate level of description Natural vs. Artificial. These results are comparable and in some cases better than the state of art approaches working on basic and superordinate level description of scenes [3, 4, 5]. For example in [3] the authors considered 13 basic classes obtaining 65,2% classification rate. We applied our technique to the same 13 classes achieving a classification rate of 67,1%.

	Suburban	Store	Buildings	Highway	Mountains	Open Country	Coast	Forest	Office	House	Overall
1	98,29	75,83	63,34	88,03	49,06	76,68	58,95	93,14	76,03	71,37	75,07
2	98,86	85,18	76,04	92,02	77,42	87,19	87,64	97,85	91,91	80,50	87,46

Table 4. Rank statistics of the two best choices.

Another way to measure the performances of the proposed approach is to use the rank statistics of the confusion matrix results which shows that the probability of a test scene correctly belongs to one of the most probable categories (Table 4). Using the two best choices, the mean categorization result increases to 87,4%. Taking into account the rank statistics, it is straightforward to show that most of the images which are incorrectly categorized as first match are on the borderline between two similar categories and therefore most often correctly categorized with the second best match (e.g., Office is classified as House).

We finally performed experiments to compare the performances of the classic bag of visual words model (corresponding to the level 0 in the hierarchy of Figure 1), with respect to the proposed hierarchical representation; experiments have

shown that the proposed hierarchy model achieves better results (8% on average) with respect to the classic bag of visual words model.

4. CONCLUSION AND PERSPECTIVE

This paper has presented an approach for scene categorization based on bag of visual words representation. The classical approach is augmented by computing it on subregions defined by three different hierarchically subdivision schemes and using a weighted distance between images. The proposed method has shown promising results despite to the fact that no complex learning procedure has been used.

Future research will be devoted to address the categorization problem on consumer digital cameras domain.

5. REFERENCES

- [1] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *Int. Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, April 2007.
- [2] S. Battiato, G. M. Farinella, G. Giuffrida, C. Sismeiro, and G. Tribulato, "Combining visual and text features for learning in multimedia direct marketing domain," in *VISAPP 2008, Int. Conf. on Comp. Vision Theory and Applications*, 2008.
- [3] Li Fei-Fei and Pietro Perona, "A hierarchical bayesian model for learning natural scene categories," in *IEEE Int. Conf. of Computer Vision and Pattern Recognition*, San Diego, CA, USA, June 2005.
- [4] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Int. Journal of Computer Vision*, vol. 42, pp. 145–175, 2001.
- [5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006, vol. II, pp. 2169–2178.
- [6] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE Int. Conf. on Computer Vision*, Oct. 2003, vol. 2, pp. 1470–1477.
- [7] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, pp. 91–97, 1981.
- [8] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–575, 2003.