Spatial Hierarchy of Textons Distributions for Scene Classification

S. Battiato[1], G. M. Farinella[1], G. Gallo[1], and D. Ravì[1]

Image Processing Laboratory, University of Catania, IT
{battiato, gfarinella, gallo, ravi}@dmi.unict.it

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 subregion the Textons distribution and the extension of the corresponding subregion are taken into account. The bags of visual words computed on the subregions are weighted and used to represent the whole scene. The classification of scenes is carried out by a Support Vector Machine. A k-nearest neighbor algorithm and a similarity measure based on Bhattacharyya coefficient are used to retrieve from the scene database those that contain similar visual content to a given a scene used as query. Experimental tests using fifteen different scene categories show that the proposed approach achieves good performances with respect to the state of the art methods.

Key words: Scene Classification, Textons, Spatial Distributions

1 Introduction

The automatic recognition of the context of a scene is a useful task for many relevant computer vision applications, such as object detection and recognition [1], content-based image retrieval (CBIR) [2] or bootstrap learning to select the advertising to be sent by Multimedia Messaging Service (MMS)[3]. Existing methods works extracting local concepts directly on spatial domain [4–6, 2] or in frequency domain [7, 8]. 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 local features [9] as well as metadata information collected during acquisition time [10] have been exploited to improve the classification task. Typically, memory-based recognition algorithms (e.g., k-nearest neighbor 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 [8].

In this paper we propose to recognize scene categories by means of bags of visual words [11] computed after hierarchically partitioning the images in subregions. Specifically, each subregion is represented as a distribution of Textons [12, 6, 13]. A weight inversely proportional to the extension of the related subregion is assigned to every distribution. The weighted Textons distributions are concatenated to compose the final representation of the scene. Like in [9] we penalize distributions related to larger regions because they can involve increasingly dissimilar visual words. The scene classification is achieved by using a SVM [14].

The proposed approach has been experimentally tested on a database of about 4000 images belonging to fifteen different basic categories of scene. In spite of the simplicity of the proposal, the results are promising: the classification accuracy obtained closely matches the results of other state-of-the-art solutions [5, 9, 8]. To perform a visual assessment, the proposed representation of the scene is used in a simple content based retrieval system employing a k-nearest neighbor as engine and a similarity measure based on Bhattacharyya coefficient [15].

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/or other visual cues such as corners, edges or local interest points. These information can be automatically extracted using several 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) task.

Local features denote distinctive patterns encoding properties of the region from which have been generated. In Computer Vision these patterns are usually referred as “visual words” [4, 16, 17, 9, 11, 13]: an image may hence be considered as a bag of “visual 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 a “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.

This type of approach leaves out the information about the spatial layout of the local features [9]. Differently than in text documents domain, the 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, the visual concepts “sky” and “sea” could be similar in terms of local descriptor, but are typically different in terms of position within the scene. The relative position can be thought as the context in which a visual word takes part respect to the other visual words within an image.

To overcome these difficulties we augment the basic bag of visual words representation combining it with a hierarchical partitioning of the image. More precisely, we partition an image using three different modalities: horizontal, vertical and regular grid. These schemes are recursively applied to obtain a hierarchy of subregions as shown in Figure 1.

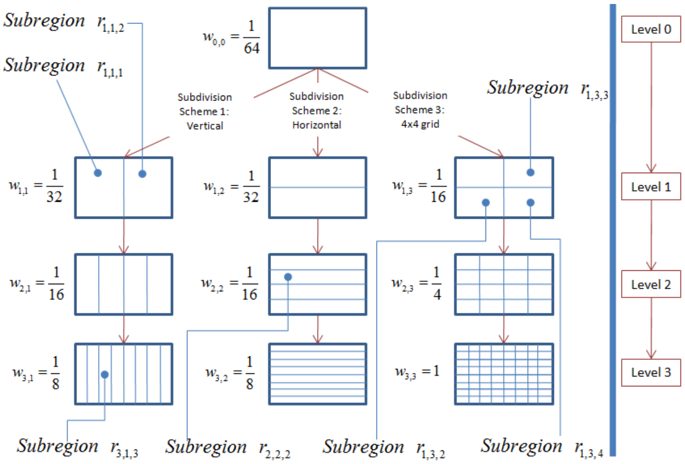


Fig. 1. Subdivision schemes up to the fourth hierarchical levels. The i_{th} subregion at level l in the subdivision scheme s is identified by $r_{l,s,i}$. The weights $w_{l,s}$ are defined by the Equation (1).

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 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).

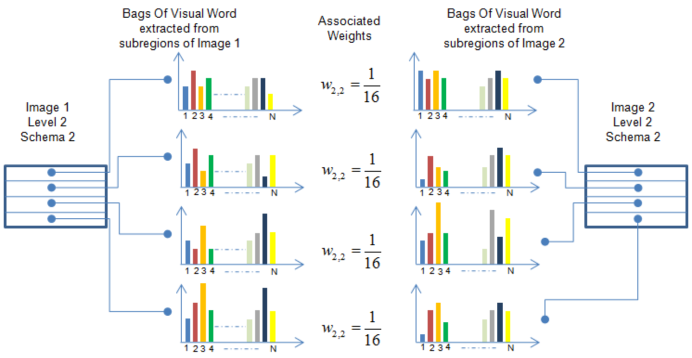


Fig. 2. A toy example of the similarity evaluation between two images I_1 and I_2 at level 2 of the subdivision schema 2. After representing each subregion $r^I_{2,2,i}$ as a distribution of Textons $B(r^I_{2,2,i})$, the distance $D_{2,2}(I_1, I_2)$ between the two images is computed taking into account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

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:

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

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

The similarity measure on the weighted bags of Textons scheme is coupled with a k-nearest neighbor algorithm to retrieve, from the scene database, those that contain similar visual content to a given scene query. To recognize the category of a scene, the weighted bags of Textons of all subregions are concatenated to form a global feature vector. The classification is obtained using a Support Vector Machine [14].

In the following subsections we provide more details about the local features used to build the bag of visual words representation as well as more details on the the similarity between images.

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 [8]. Because humans can process texture quickly and in parallel over the visual field, we considered texture as a good holistic cue candidate. Specifically, we choose to use Textons [12, 6, 13] 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.

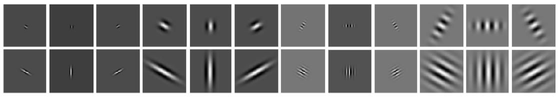


Fig. 3. Visual representation of the bank of 2D Gabor filters used in our experiments.

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 [18]:

$$G(x, y, f_0, \theta, \alpha, \beta) = e^{-(\alpha^2 x_2^{'} + \beta^2 y_2^{'})} \times e^{j2\pi f_0 x^{'}} \tag{2}$$

$$x^{'} = x \cos\theta + y \sin\theta \tag{3}$$

$$y^{'} = -x \sin\theta + y \cos\theta \tag{4}$$

The 24 Gabor filters (Figure 3) 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 (θ = -60°, 0, 60°), two different sharpness of the Gaussian major axis (α = 0.5, 1.5) and two different sharpness of the Gaussian minor axis (β = 0.5, 1.5). Each filter is centered at the origin and no phase-shift is applied.

The experiments reported in Section 3 are performed by using a spatial hierarchy with three level (l = 0,1,2) and employing a visual vocabulary of 400 Textons.

2.2 Similarity Between Images

The weigthed distance that we use is founded on similarity between two corresponding subregions when the bag of visual words have been computed on the same vocabulary.

Let B($r^{I_1}_{l,s,i}$) and B($r^{I_2}_{l,s,i}$) 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($r^{I_1}_{l,s,i}$) and B($r^{I_2}_{l,s,i}$). Such distance measure has several desirable properties [15]: 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 computed as follows:

$$D_{l,s}(I_1, I_2) = w_{l,s} * \sum_{i} \sqrt{1 - \rho[B(r^{I_1}_{l,s,i}), B(r^{I_2}_{l,s,i})]} \tag{5}$$

$$\rho[B(r^{I_1}_{l,s,i}), B(r^{I_2}_{l,s,i})] = \sum_{T} \sqrt{B(r^{I_1}_{l,s,i})_T * B(r^{I_2}_{l,s,i})_T} \tag{6}$$

where $B(r^{I}_{l,s,i})_T$ indicate the frequency of a specific Texton T in the subregion $r_{l,s,i}$ of the image I. The final distance between two images I_1 and I_2 is hence calculated as follows:

$$D(I_1, I_2) = D_{0,0} + \sum_{l} \sum_{s} D_{l,s} \tag{7}$$

Observe that the level $l = 0$ of the hierarchy (Figure 1) corresponds to the classic bag of visual word model in which the metric based on Bhattacharyya coefficient is used to establish the distance between two images.

3 Experiments and Results

The dataset we have used contains more than 4000 images collected in [5, 9, 8]. Images are grouped in fifteen basic categories of scenes (Figure 4): coast, forest, bedroom, kitchen, living room, suburban, office, open countries, mountains, tall building, store, industrial, inside city, highway. These basic categories can be ensembled and described with a major level of abstraction (Figure 4): In vs. Out, Natural vs. Artificial. Moreover, some basic categories (e.g., bedroom, living room, kitchen) can be grouped and considered belonging to a single category (e.g. house).

In our experiments we splitted the database in ten different non overlapped subsets. Each subset was created in order to have approximatively the 10% of images of a specific class. The classification experiments have been repeated ten times considering the i_{th} subset as training and the remaining subsets as test. A ν-SVC [19] was trained at each run and the per-class classification rates were recorded in a confusion matrix in order to evaluate the classification performance 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 79.43% considering the fifteen basic classes, 97.48% considering the superordinate level of description Natural vs. Artificial, 94.5% considering the superordinate level of description In vs. Out. These results are comparable and in some cases better than the state of art approaches working on basic and superordinate level description of scenes [5, 7, 9, 8, 20]. For example, in [5] the authors considered thirteen basic classes obtaining 65.2% classification rate. We applied the proposed technique to the same dataset used in [5] achieving a classification rate of 84% (Figure 5).

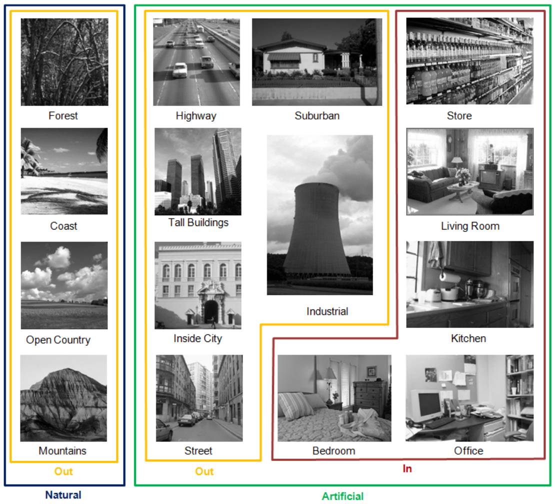
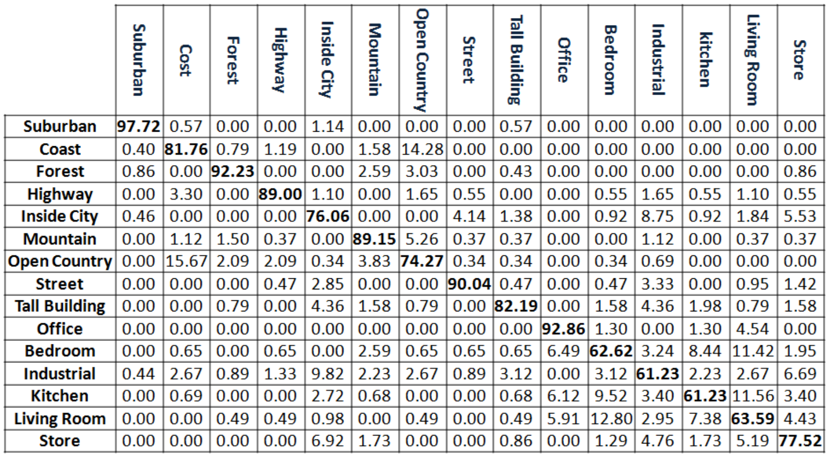
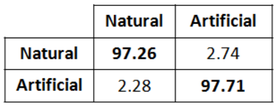


Fig. 4. Some examples of images used in our experiments considering basic and superordinate level of description.



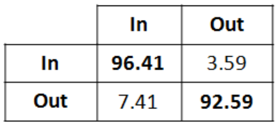
	Suburban	Cost	Forest	Highway	Inside City	Mountain	Open Country	Street	Tall Building	Office	Bedroom	Industrial	kitchen	Living Room	Store
Suburban	**97.72**	0.57	0.00	0.00	1.14	0.00	0.00	0.00	0.57	0.00	0.00	0.00	0.00	0.00	0.00
Coast	0.40	**81.76**	0.79	1.19	0.00	1.58	14.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Forest	0.86	0.00	**92.23**	0.00	0.00	2.59	3.03	0.00	0.43	0.00	0.00	0.00	0.00	0.00	0.86
Highway	0.00	3.30	0.00	**89.00**	1.10	0.00	1.65	0.55	0.00	0.00	0.55	1.65	0.55	1.10	0.55
Inside City	0.46	0.00	0.00	0.00	**76.06**	0.00	0.00	4.14	1.38	0.00	0.92	8.75	0.92	1.84	5.53
Mountain	0.00	1.12	1.50	0.37	0.00	**89.15**	5.26	0.37	0.37	0.00	0.00	1.12	0.00	0.37	0.37
Open Country	0.00	15.67	2.09	2.09	0.34	3.83	**74.27**	0.34	0.34	0.00	0.34	0.69	0.00	0.00	0.00
Street	0.00	0.00	0.00	0.47	2.85	0.00	0.00	**90.04**	0.47	0.00	0.47	3.33	0.00	0.95	1.42
Tall Building	0.00	0.00	0.79	0.00	4.36	1.58	0.79	0.00	**82.19**	0.00	1.58	4.36	1.98	0.79	1.58
Office	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**92.86**	1.30	0.00	1.30	4.54	0.00
Bedroom	0.00	0.65	0.00	0.65	0.00	2.59	0.65	0.65	0.65	6.49	**62.62**	3.24	8.44	11.42	1.95
Industrial	0.44	2.67	0.89	1.33	9.82	2.23	2.67	0.89	3.12	0.00	3.12	**61.23**	2.23	2.67	6.69
Kitchen	0.00	0.69	0.00	0.00	2.72	0.68	0.00	0.00	0.68	6.12	9.52	3.40	**61.23**	11.56	3.40
Living Room	0.00	0.00	0.49	0.49	0.98	0.00	0.49	0.00	0.49	5.91	12.80	2.95	7.38	**63.59**	4.43
Store	0.00	0.00	0.00	0.00	6.92	1.73	0.00	0.00	0.86	0.00	1.29	4.76	1.73	5.19	**77.52**

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



	Natural	Artificial
Natural	**97.26**	2.74
Artificial	2.28	**97.71**

Table 2. Natural vs. Artificial results.



	In	Out
In	**96.41**	3.59
Out	7.41	**92.59**

Table 3. In vs. Out results.

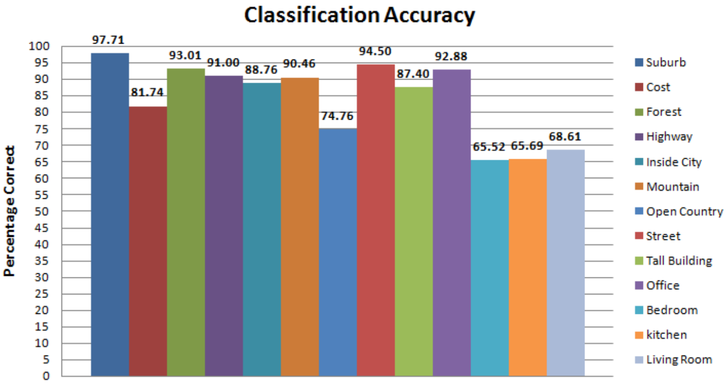


Fig. 5. Classification accuracy considering the thirteen basic categories used in [5].

Obviously, the classification accuracy increases (≅89%) if the images belonging to the categories bedroom, kitchen and living room are grouped and described as house scene. Moreover, the method proposed in this paper achieves better results with respect to our previous work [20], were the overall classification rate was 75% considering the only 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.

Another way to measure the performances of the proposed approach is to use the rank statistics [5, 2] of the confusion matrix results. Rank statistics shows the probability of a test scene correctly belongs to one of the most probable categories (Table 4). Using the two best choices on the fifteen basic classes, the mean categorization result increases to 86.99% (Table 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., Coast is classified as Open Country).
Finally we compared 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 model achieves better results (8% on average).

To perform a visual assessment, Figure 6 shows some examples of images retrieved employing a k-nearest neighbor and the similarity measure described in Section 2. The query images are depicted in the first column, whereas the first three closest images are reported in the other columns. The closest images are semantically consistent in terms of visual content to the related query images.

	Suburban	Cost	Forest	Highway	Inside City	Mountain	Open Country	Street	Tall Building	Office	Bedroom	Industrial	kitchen	Living Room	Store	Overall
1	97.72	81.76	92.23	89.00	76.06	89.15	74.27	90.04	82.19	92.86	62.62	61.23	61.23	63.59	77.52	**79.43**
2	98.29	96.04	95.26	92.30	84.81	94.41	89.94	93.37	86.55	97.40	74.04	71.05	72.79	76.39	82.28	**86.99**

Table 4. Rank statistics of the two best choices on the fifteen basic classes.

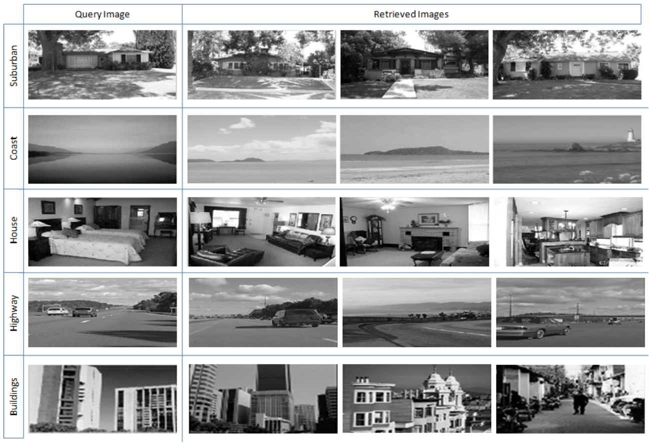


Fig. 6. Examples of images retrieved employing the proposed approach. The query images are on the left, and top three closest images are shown on the right.

4 Conclusion

This paper has presented an approach for scene categorization based on bag of visual words representation. The classic approach is augmented by computing it on subregions defined by three different hierarchically subdivision schemes and properly weighting the Textons distributions with respect to the involved subregions. The weighed bags of visual words representation is coupled with a Support Vector Machine to perform classification. A similarity distance based on Bhattacharyya coefficient is used together with a k-nearest neighboor to retrieve scenes. Despite its simplicity, the proposed method has shown promising results with respect to state of the art methods. The proposed hierarchy of features produces a description of the image only slightly heavier than the classical bag of words representation, both in terms of storage as well as in terms of time retrieval allowing at the same time to obtain effective results. Future works should be devoted to perform a depth comparison between different kind of features used to build the visual vocabulary (e.g., Textons vs. SIFT) for scene classification. Moreover, the proposed method should be compared with respect to other approaches working on spatial (e.g., Spatial Pyramid Matching [9], pLSA [4], etc.) as well as on frequency domains [21, 22].

References

1. Torralba, A.: Contextual priming for object detection. International Journal of Computer Vision **53**(2) (2003) 169–191
2. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision **72**(2) (April 2007) 133–157
3. Battiato, S., Farinella, G.M., Giuffrida, G., Sismeiro, C., Tribulato, G.: Using visual and text features for direct marketing on multimedia messaging services domain. Multimedia Tools and Applications Journal (In Press) (2008)
4. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: Proceedings of the European Conference on Computer Vision. (2006)
5. Fei-Fei, L., Perona, P.: A hierarchical bayesian model for learning natural scene categories. In: IEEE Computer Science Society International Conference of Computer Vision and Pattern Recognition, (CVPR), San Diego, CA, USA (June 2005)
6. Renninger, L.W., Malik, J.: When is scene recognition just texture recognition? Vision Research **44** (2004) 2301–2311
7. Ladret, P., Guérin-Dugué, A.: Categorisation and retrieval of scene photographs from jpeg compressed database. Pattern Analysis & Application **4** (June 2001) 185–199
8. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision **42** (2001) 145–175
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume II. (2006) 2169–2178
10. Matthew, R.B., Jiebo, L.: Beyond pixels: Exploiting camera metadata for photo classification. Pattern Recognition **38**(6) (2005) 935–946
11. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision. Volume 2. (October 2003) 1470–1477
12. Julesz, B.: Textons, the elements of texture perception, and their interactions. Nature **290** (1981) 91–97
13. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. International Journal of Computer Vision **62**(1–2) (April 2005) 61–81
14. Shawe-Taylor, J., Cristianini, N.: Support Vector Machines and other kernel-based learning methods. Cambridge University Press (2000)
15. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5) (2003) 564–575
16. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision. (2004)

17. J. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: IEEE Computer Science Society International Conference of Computer Vision and Pattern Recognition, (CVPR). (2008)
18. Gonzalez, R.C., Woods, R.E.: Digital Image Processing (3rd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA (2006)
19. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001)
20. Battiato, S., Farinella, G.M., Gallo, G., Ravì, D.: Scene categorization using bag of textons on spatial hierarchy. In: IEEE International Conference on Image Processing - ICIP 2008. (2008)
21. Farinella, G.M., Battiato, S., Gallo, G., Cipolla, R.: Natural Versus Artificial Scene Classification by Ordering Discrete Fourier Power Spectra. In: To appear in Proceedings of 12th International Workshop on Structural and Syntactic Pattern Recognition, (SSPR)- Satellite event of the 19th International Conference of Pattern Recognition (ICPR) - Lecture Notes in Computer Science. (2008)
22. Battiato, S., Farinella, G.M., Gallo, G., Messina, E.: Classification of compressed images in constrained application domains. In: SPIE-IS&T 21th Annual Symposium Electronic Imaging Science and Technology 2009 - Digital Photography V. (2009)