Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2010, Article ID 919367, 13 pages
doi:10.1155/2010/919367

Research Article

Exploiting Textons Distributions on Spatial Hierarchy for Scene Classification

S. Battiato, G. M. Farinella, G. Gallo, and D. Ravì

Image Processing Laboratory, University of Catania, 95125 Catania, Italy

Correspondence should be addressed to G. M. Farinella, gfarinella@dmi.unict.it

Received 29 April 2009; Revised 24 November 2009; Accepted 10 March 2010

Academic Editor: Benoit Huet

This paper proposes a method to recognize scene categories using bags of visual words obtained by 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 discriminative methods (i.e., SVM, KNN). A similarity measure based on Bhattacharyya coefficient is proposed to establish similarities between images, represented as hierarchy of bags of visual words. Experimental tests, using fifteen different scene categories, show that the proposed approach achieves good performances with respect to the state-of-the-art methods.

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, 4].

Existing methods work on extracting local concepts directly on spatial domain [2, 5–7] or frequency domain [8, 9]. A global representation of the scene is obtained by grouping together local information in different ways (e.g., histogram of visual concepts, spectra templates, etc.). Recently, the spatial layout of local features [10–13] as well as metadata information collected during acquisition time [14] have been exploited to improve the classification task. Typically, memory-based recognition algorithms (e.g., Support Vector Machine [15], K-nearest neighbors [16]), etc.) are employed, together with holistic representation, to classify scenes skipping the recognition of the objects within the scene [9].

In this paper, we propose to recognize scene categories by means of bags of visual words [17] computed after hierarchically partitioning the images in subregions. Specifically, each subregion is represented as a distribution of Textons [7, 18, 19]. 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 [10], we penalize distributions related to larger regions because they can involve increasingly dissimilar visual words. The scene classification is achieved by using a discriminative method (i.e., SVM or KNN). Differently than [10–13], we use Textons rather than SIFT based features [20] and an augmented spatial pyramid involving together three subdivision schemes: horizontal, vertical, and regular grid. Also we use a linear kernel (rather than a pyramid one) during SVM classification, whereas a similarity measure based on Bhattacharyya coefficient [21] (instead of χ^2 distance) when KNN is employed for classification purpose.

To allow a straightforward comparison with state-of-the-art methods [6, 8–10] the proposed approach has been experimentally tested on a benchmark 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 [6, 8–10].

The rest of the paper is organized as follows: Section 2 briefly reviews related works in the field. Section 3 describes the model we have used to represent images. Section 4 illustrates the dataset, the setup involved in our experiments, and the results obtained using the proposed approach. Finally, in Section 5 we conclude with avenues for further research.

2. Related Works

Scene understanding is a fundamental process of human vision that allows us to efficiently and rapidly analyze our surroundings. Humans are able to recognize complex visual scenes at a single glance, despite the number of objects with different poses, colors, shadows, and textures that may be contained in the scenes. Understanding the robustness and rapidness of this human ability has been a focus of investigation in the cognitive sciences over many years. Seminal studies in computational vision [22] have portrayed scene recognition as a progressive reconstruction of the input from local measurements (e.g., edges, surfaces). In contrast, some experimental studies have suggested that recognition of real world scenes may be initiated from the encoding of the global configuration, ignoring most of the details about local concepts, and object information [23]. This ability is achieved mainly by exploiting the holistic cues of scenes that can be processed as single entity over the entire human visual field without requiring attention to local features [24, 25].

The advancements in image understanding have inspired computer vision researchers to develop computational systems capable of automatically recognizing the category of scenes. The recognition of the context of a scene is a useful task for many relevant computer vision applications:

(i) context driven focus attention, object priming, and scale selection [1];

(ii) content-based image retrieval (CBIR) [2];

(iii) semantic organization of databases of digital pictures [8];

(iv) robot navigation systems [26];

(v) scene depths estimation [27, 28];

(vi) bootstrap learning to select the best advertising to be sent by Multimedia Messaging Service [3, 4].

Recent studies suggested that humans rely on local information as much as on global information to recognize the scene category. Specifically, the Human Visual System seems to integrate both type of information during the categorization of scenes [29].

In building scene recognition systems some consideration about the spatial envelope properties (e.g., degree of naturalness, degree of openness, etc.) and the level of description (e.g., subordinate, basic, superordinate) of the scenes should be taken into account [9]. Levels of description that use precise semantic names to categorize an environment (e.g., beach, street, forest) do not explicitly refer to the scene structure. Hence, the spatial envelop of a scene should be taken into account and encoded in the scene representation model independently from the required level of scene description. Moreover, the scene representation model and the related computational approach depend on the task to be solved and the level of description required.

Different methods have been proposed to model the scene in order to build an expressive description of the content. Existing methods work on extracting local concepts directly on spatial domain [2, 6, 7] or frequency domain [9, 30]. A global representation of the scene may be obtained by grouping together these information in different ways. Recently, the spatial layout of the local information [10–13, 31] as well as metadata information collected during acquisition time [14] have been used to improve the classification accuracy.

The final descriptor of the scene is eventually exploited by some pattern recognition algorithms to infer the scene category, skipping the recognition of the objects that are present in the scene [9]. Machine learning procedures are employed to automatically learn commonalities and differences between different classes.

In the following, we will illustrate in more details some of the state-of-the-art approaches working with features extracted on spatial domain.

2.1. Scene Classification Extracting Local Concepts on Spatial Domain. Several studies in Computer Vision have considered the problem of discriminating between classes at superordinate level of description. A wide class of scene recognition algorithms use color, texture, or edge features. Gorkani and Picard used statistics of orientation in the images to discriminate a scene into two categories (*cities* and *natural landscapes*) [32]. *Indoor* versus *Outdoor* classification based on color and texture was addressed by Szummer and Picard [33]. Many other authors proposed to organize and classify images by using the visual content encoded on spatial domain. In this section, we review some existing works for scene classification focusing on methods that use features extracted on spatial domain. Other related approaches are reviewed in [34].

Renninger and Malik employed a holistic representation of the scene to recognize its category [7]. The rationale in using a holistic representation was that the holistic cues are processed over the entire human visual field and do not require attention to analyze local features, allowing humans to recognize quickly the category of the scene. Taking into account that humans can process texture quickly and in parallel over the visual field, a global representation based on Textons [18, 24] has been used. The main process used to encode textures started by building a vocabulary of distinctive patterns, able to identify properties and structures of different textures present in the scenes. The vocabulary was built using K-means clustering on a set of filter responses. Using the built vocabulary, each image is represented as a frequency histogram of Textons. Images

of scenes used in the experiments were within ten basic-level categories: *beach*, *mountain*, *forest*, *city*, *farm*, *street*, *bathroom*, *bedroom*, *kitchen,* and *living room*. A χ^2 similarity measure was coupled with a K-nearest neighbors algorithm to perform classification. The performances of the proposed model stayed nearly at 76% correct.

Fei-Fei and Perona suggested an approach to learn and recognize natural scene categories with the interesting peculiarity that it does not require any experts to annotate the training set [6]. The dataset involved in their experiments contained thirteen basic level categories of scenes: *highway*, *inside of cities*, *tall buildings*, *streets*, *forest*, *coast*, *mountain*, *open country*, *suburb residence*, *bedroom*, *kitchen*, *living room,* and *office*. The images of scenes were modeled as a collection of local patches automatically detected on scale invariant points and described by a features vector invariant to rotation, illumination, and 3D viewpoint [20]. Each patch was represented by a codeword from a large vocabulary of codewords previously learned through K-means clustering on a set of training patches. In the learning phase a model that represents the best distribution of the involved codewords in each category of scenes was built by using a learning algorithm based on Latent Dirichlet Allocation [35]. In recognition phase, first the identification of all the codewords in the unknown image was done. Then the category model that best fitted the distribution of the codewords of a test image was inferred comparing the likelihood of an image given each category. The performances obtained by authors reaches 65.2% of accuracy.

The goal addressed by Bosch et al. in [5] was to discover the objects in each image in an unsupervised manner, and to use the distribution of objects to perform scene classification. To this aim, probabilistic Latent Semantic Analysis (pLSA) [36] was applied to a bag of visual words representation of each image. A new visual vocabulary for the bag of visual word model exploiting the SIFT descriptor [20] on HSV colour domain has been proposed. As usual, K-means was employed to build the vocabulary. The scene classification on the object distribution was carried out by a K-nearest neighbors classifier. The combination of (unsupervised) pLSA followed by (supervised) nearest neighbors classification proposed in [5] outperformed previous methods. For instance, the accuracy of this approach was 8.2% better with respect to the method proposed in [6] when compared on the same dataset.

One of the most complete scene category dataset at basic level of description was exploited by Lazebnik et al. [10]. The dataset they used is an augmented version of the dataset used in [5, 6] in which two basic level categories have been added: *industrial* and *store*. The proposed method exploits a spatial pyramid image representation building on the idea proposed in [37] in which a *Pyramid Match Kernel* is used to find an approximate correspondence between two sets of elements. For a kind of visual words (e.g., corner [38], SIFT [20], etc.), it first identifies where spatially the visual word appears in the image. Then at each level of the pyramid, the subimages of the previous level are splitted in four subimages. A histogram for each subimage in the pyramid is built containing for each bin the frequency of a specific visual word. Finally, the spatial pyramid image representation is obtained as the vector containing all histograms weighted taking into account the corresponding level. The weights associated to each histogram are used to penalize the match of two corresponding histogram bins related to a larger subimage and emphasizes match when bins refer to a smaller subimage. The authors employed a SVM using the one-versus-all rule to perform the recognition of the scene category. This method obtained 81.4% when SIFT descriptors of 16×16 pixels patches computed over a grid with 8 pixels spacing were employed in building the visual vocabulary through K-means clustering. Although the spatial hierarchy we propose in this paper in some sense resembles the work in [10], it introduces a different scheme of splitting the image in the hierarchy, a different way to weight the contribution of each subregion, as well as a different similarity criterion between histograms.

Vogel and Schiele considered the problem of identifying natural scenes within six different basic level categories [2]. The basic involved category in the experiments was related to *costs*, *rivers/lakes*, *forests*, *plains*, *mountains*, and *sky/clouds*. A novel image representation was introduced. The scene model takes into account nine local concepts that can be present in *Natural* scenes (*sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers*, *sand*) and combines them to a global representation used to address the category of the scenes. The descriptor for each image scene is built in two stages. First, local image regions are classified by a concept classifier taking into account the nine semantic concept classes. The region-wise information of the concept classifier is then combined to a global representation through a normalized vector in which each component represents the frequency of occurrence of a specific concept taking into account the image labeled in the first stage. In order to model information about which concept appears at any specific part of the image (e.g., top, bottom), the vector of frequency concepts was computed on several overlapping or nonoverlapping image areas. In this manner a semilocal spatial image representation by computing and concatenating the different frequency vectors can be obtained. To perform concept classification each concept patch of an image was represented by using low level features (HIS color histogram, edge directions histogram, and gray-level co-occurrence). A multiclass SVM using a one-against-one approach was used to infer local concepts as well as the final category of the scene. The best classification accuracy obtained with this approach was 71,7% for the nine concepts and 86,4% for the six classes of scene.

Recently, Bosch et al., inspired by previous works [5, 10], presented a method in which the pLSA model was augmented using spatial pyramid in building the distribution of latent topics [39]. The final scene classification was performed using the discriminative classifier SVM on the learned distribution obtaining 83.7% of accuracy on the same dataset used in [10].

In sum, all of the approaches above share the same basic structure that can be schematically summarized as follows.

(1) A suitable features space is built (e.g., visual words vocabulary). The space emphasizes specific image cues such as, for example, corners, oriented edges, textures, and so forth.

(2) Each image is projected into this space. A descriptor, as a whole entity, of the image projection in the feature space is built (e.g., visual words histograms).

(3) Scene classification is obtained by using pattern recognition and machine learning algorithms on the holistic representation of the images.

A wide class of classification algorithms based on the above scheme work on extracting features on perceptually uniform color spaces (e.g., CIELab). Typically, filter banks or local invariant descriptors are employed to capture image cues and to build the visual vocabulary to be used in a bag of visual words model. An image is considered as a distribution of visual words and this holistic representation is used to perform classification. Eventually, local spatial constraints are added in order to capture the spatial layout of the visual words within images [2, 10].

Recent works [11–13] demonstrated that augmenting the spatial pyramid image representation proposed in [10] through a horizontal subdivision scheme is useful to improve the recognition accuracy when SIFT-based descriptors are employed as local features. In this paper, we propose a new framework involving together three different subdivision schemes to build a hierarchy of bags of Textons.

3. 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 by grouping together such local information. This representation is then used during categorization task. Local features denote distinctive patterns encoding properties of the region from which they have been generated. In Computer Vision these patterns are usually referred to as "visual words" [5, 10, 17, 40]: 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 can be properly 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 [10–13]. 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. Despite spatial pyramid with different subdivision schemes have been already adopted [10–13], the three subdivision schemes proposed here have been never used together before. Experiments confirm the effectiveness of such strategy as reported by the measured performances reported into the experimental section.

The bag of visual words representation is hence computed in the usual way on each subregion, using a set of prebuilt vocabularies corresponding to different levels in the hierarchy. Specifically, for each level of the hierarchy a corresponding vocabulary is built and used. In our experiments we use Textons as visual words. The proposed augmented representation hence, keeps record of the frequency of Textons in each subregion (Figure 2), taking into account the vocabulary corresponding to the level under consideration. In this way we take into account the spatial layout information of local features.

A similarity measure between images is defined as follows. First, a similarity measure between histograms of visual words relative to corresponding regions is computed (as reported in Section 3.2). The connection of similarity values of each subregion are then combined into a final distance by means of a weighted sum. The choice of weights 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 (Figures 1 and 2).

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

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

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

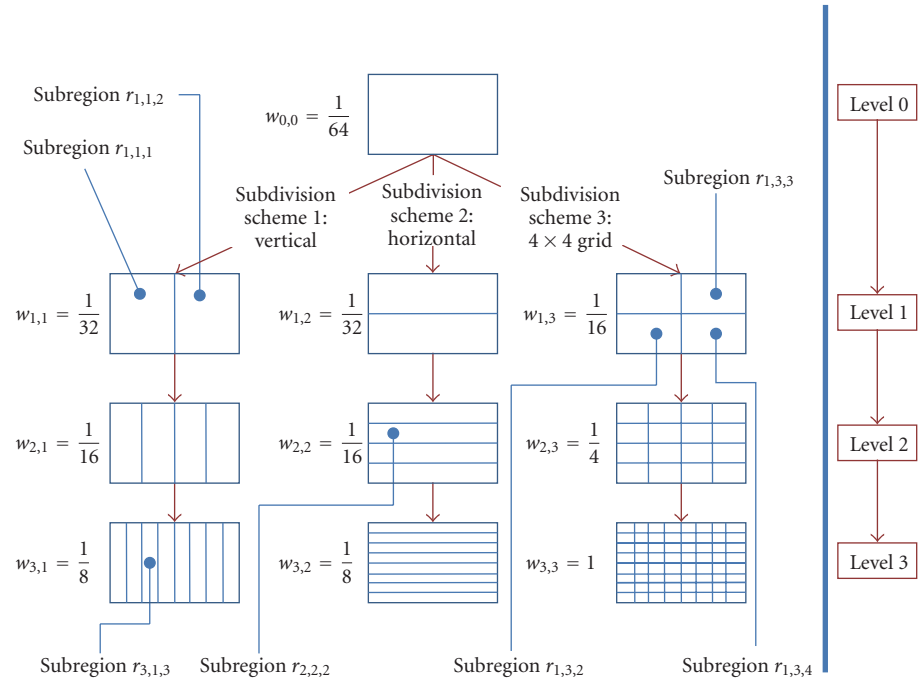

FIGURE 1: Subdivision schemes up to the fourth hierarchical levels. The ith subregion at level l in the subdivision scheme s is identified by $r_{l,s,i}$. The weights $w_{l,s}$ are defined by (1).

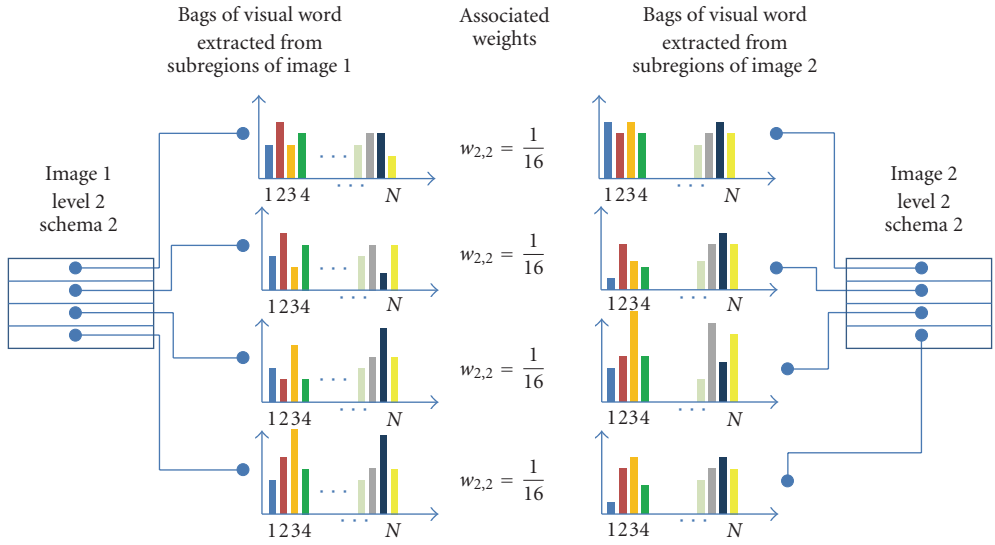

FIGURE 2: A toy example of the similarity evaluation between two images I_1 and I_2 at level 2 of the subdivision schema 2. After representing each subregion $r^l_{2,2,i}$ as a distribution of Textons $B(r^l_{2,2,i})$, the distance $D_{2,2}(I_1, I_2)$ between the two images is computed taking into account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

The similarity measure on the weighted bags of Textons scheme can be used with a K-nearest neighbors algorithm for classification purposes. In performing categorization with SVM, the weighted bags of Textons of all subregions are concatenated to form a global feature vector.

Considering a hierarchy with L levels and a visual vocabulary V_l with T_l Textons at level l, the feature vector associated to an image has dimensionality $T_0 + \sum_{l=1}^{L} T_l(2^{l+1} + 4^l)$.

In the experiments reported in Section 4, effective results have been obtained by considering $L = 2$, and the vocabularies V_0, V_1, V_2 with, respectively, $T_0 = 400$, $T_1 = 200$, and $T_2 = 100$ Textons, resulting in a 4400 dimensional feature vector containing the histograms of all subregions involved in the considered hierarchy. We have used integral histograms [41] to reduce both the space used to store an image represented as bags of Textons, and the time

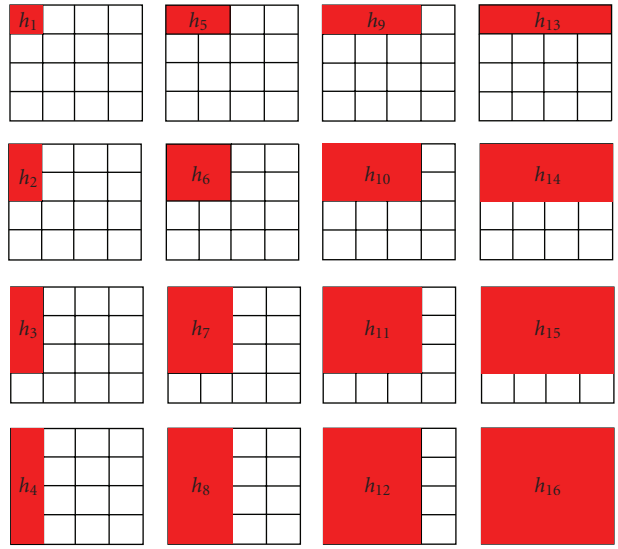

Figure 3: Example of integral histogram representation used at level $l = 2$ of the scheme 3. The ith subregion level $l = 2$ of the scheme 3 in Figure 1 is associated to a histogram h_i computed on the red area taking into account the vocabulary with T_2 Textons.

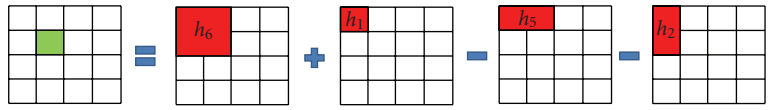

Figure 4: Histograms related to subregions in the hierarchy are computed exploiting the integral histogram representations. In this example the histogram $H_{2,3,6}$ related to the subregion $r_{2,3,6}$ in the hierarchy with $L = 2$ levels is computed considering the integral histogram representation at level $l = 2$ as $H_{2,3,6} = h_6 + h_1 - h_5 - h_2$.

needed in building all the histograms involved in the hierarchy from the stored information. Specifically, to store the overall representation of an image, we use the histogram at level $l = 0$, the integral histograms at level $l = 1$ of the Scheme 3, and the integral histograms at level $l = 2$ of the Scheme 3 (Figure 3). In this way we need to store $\sum_{l=0}^{2} 4^l$ histograms resulting in a feature vector of dimensionality $\sum_{l=0}^{2} T_l 4^l = 2800$. All the histograms related to subregions in the hierarchy are computed by using basic operations on the integral histograms representations (Figure 4).

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

3.1. 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 [9]. 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 [7, 18, 19] 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, results presented in Section 4 have been obtained by considering a bank of 2D Gabor filters (In our experiments 2D Gabor filters slightly outperformed the bank of filters used in [42].) 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:

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

The 24 Gabor filters (Figure 5) 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 ($\theta = -60°, 0, 60°$), two different sharpnesses of the Gaussian major axis ($\alpha = 0.5, 1.5$), and two different sharpnesses of the Gaussian minor axis ($\beta = 0.5, 1.5$). Each filter is centered at the origin and no phase-shift is applied. Since the used filter banks respond to basic image features (e.g., edges, bars) considered at different scales and orientations, they are innately immune to most changes in an image [7, 24, 43].

Figure 5: Visual representation of the 2D Gabor filter banks used in our experiments.

3.2. Similarity between Images. The weighted 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 $B(r_{l,s,i}^{I_1})$ and $B(r_{l,s,i}^{I_2})$ be the bags of visual words representation of the ith 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 [44]: it imposes a metric structure, it has a clear geometric interpretation, it is valid for arbitrary distributions, and 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:

$$D_{l,s}(I_1, I_2) = w_{l,s} * \sum_i \sqrt{1 - \rho\left[B\left(r_{l,s,i}^{I_1}\right), B\left(r_{l,s,i}^{I_2}\right)\right]},$$
$$\rho\left[B\left(r_{l,s,i}^{I_1}\right), B\left(r_{l,s,i}^{I_2}\right)\right] = \sum_{t \in V_l} \sqrt{B(r_{l,s,i}^{I_1})_t * B(r_{l,s,i}^{I_2})_t}, \quad (3)$$

where $B(r_{l,s,i}^{I})_t$ indicate the frequency of a specific Texton t within the vocabulary V_l 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:

$$D(I_1, I_2) = D_{0,0} + \sum_l \sum_s D_{l,s}. \quad (4)$$

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.

Considering a hierarchy with L levels and a visual vocabulary V_l with T_l Textons at level l, the number of operations involved (i.e., addition, substraction, multiplication, and root square) in the computation of the similarity measure in (4) is $[(2T_0 + 2) + 1] + \sum_{l=1}^{L}[(2T_l + 2)(2^{l+1} + 4^l) + 3]$. In the experiments reported in Section 4, we used a hierarchy with $L = 2$, and vocabularies V_0, V_1, V_2 with, respectively, $T_0 = 400$, $T_1 = 200$, and $T_2 = 100$ Textons. The average computational time needed to compute the above similarity measure between two images was 1.30300 milliseconds considering a matlab implementation running on an Intel Core Duo 2.53 GHz.

The similarity measure above outperformed other similarity measures proposed in literature (e.g., χ^2 used in [7, 13]) as reported in Section 4.

4. Experiments and Results

To allow a straightforward comparison with state-of-the-art methods [6, 8–10] the proposed approach has been experimentally tested on a benchmark database of about 4000 images collected by the authors of [6, 9, 10]. Images are grouped in fifteen basic categories of scenes (Figure 6): *coast*, *forest*, *bedroom*, *kitchen*, *living room*, *suburban*, *office*, *open countries*, *mountains*, *tall building*, *store*, *industrial*, *inside city*, and *highway*. These basic categories can be ensembled and described with a major level of abstraction (Figure 6): *In* versus *Out*, *Natural* versus *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 nonoverlapped subsets. Each subset was created in order to have approximately 10% of images of a specific class. The classification experiments have been repeated ten times considering the ith subset as training and the remaining subsets as test.

A ν-SVC [45] 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 obtained employing SVM as a classifier are reported through confusion matrices in Tables 1, 2, and 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* versus *Artificial*, and 94.5% considering the superordinate level of description *In* versus *Out*.

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 taking into account different levels, as well as the impact of the different subdivision schemes involved in the hierarchy. Results are reported in Tables 4 and 5. Experiments confirm that the proposed model achieves better results (8% on average) with respect to the standard bag of visual word model (corresponding to the level 0 of the hierarchy). Considering more than two levels in the hierarchy does not improve the classification accuracy, whereas the complexity of the model increases becoming prohibitive with more than three levels.

Experiments demonstrate also that the best results in terms of overall accuracy are obtained considering all three schemes together as reported in Table 5.

Table 1: Confusion matrix obtained considering the proposed representation and SVM classifier on the fifteen basic classes of scenes. The average classification rates for individual classes are listed along the diagonal.

	Subur ban	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
Cost	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**

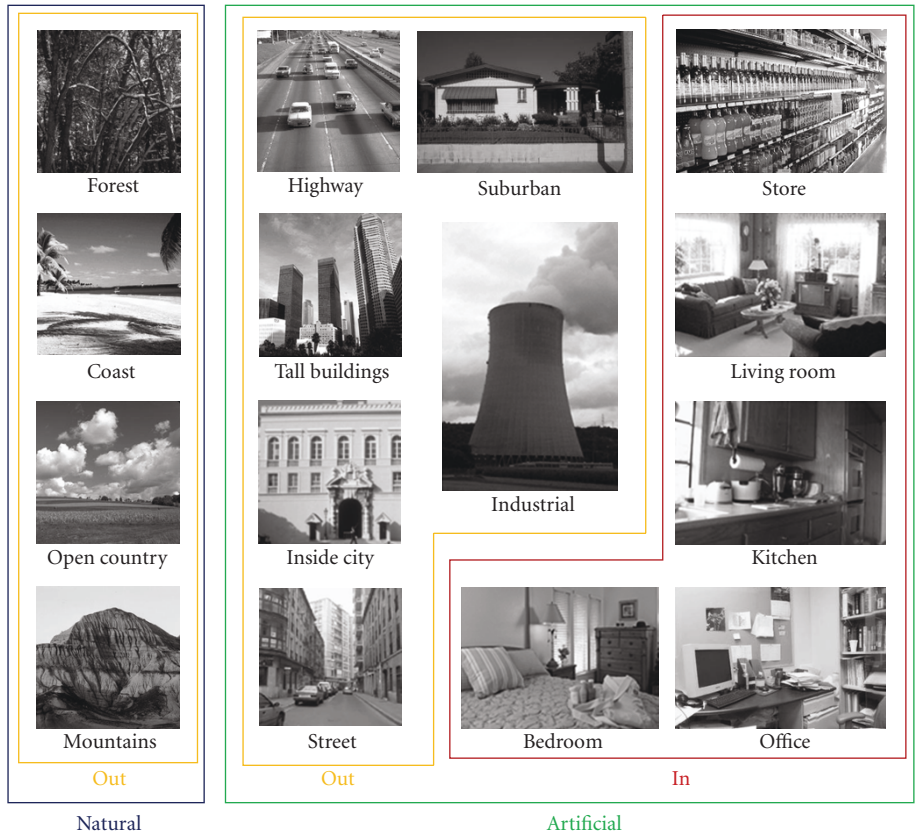

Figure 6: Some examples of images used in our experiments considering basic and superordinate levels of description.

Table 2: *Natural* versus *Artificial* results obtained considering the proposed representation and SVM classifier.

	Natural	Artificial
Natural	**97.27**	2.74
Artificial	2.28	**97.71**

Table 3: *In* versus *Out* results obtained considering the proposed representation and SVM classifier.

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

Table 4: Results obtained considering different levels in the hierarchy.

Level 0	71.39
Level 1	75.58
Level 2	79.43
Level 3	79.67

The obtained results are comparable and in some cases better than the state-of-the-art approaches working on basic and superordinate level description of scenes [6, 8–10]. For example, in [6] the authors considered thirteen basic classes obtaining 65.2% classification rate. We applied the proposed technique to the same dataset used in [6] achieving a classification rate of 84% (Figure 7). Obviously, the classification accuracy of the proposed approach increases (≅89%) if the images belonging to the categories *bedroom*, *kitchen*, and *living room* are grouped and described as *house* scene.

Table 5: Results obtained considering different schemes in the hierarchy. The best results are obtained by using the three schemes together.

Scheme	1	2	3	1 + 3	2 + 3	1 + 2 + 3
Accuracy	71.92	74.50	75.61	76.34	76.89	79.43

Another way to measure the performances of the proposed approach is to use the rank statistics [2, 6] of the confusion matrix results. Rank statistics shows the probability of a test scene to correctly belong to one of the most probable categories. Using the two best choices on the fifteen basic classes, the mean categorization result increases to 86.99% (Table 6). 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*).

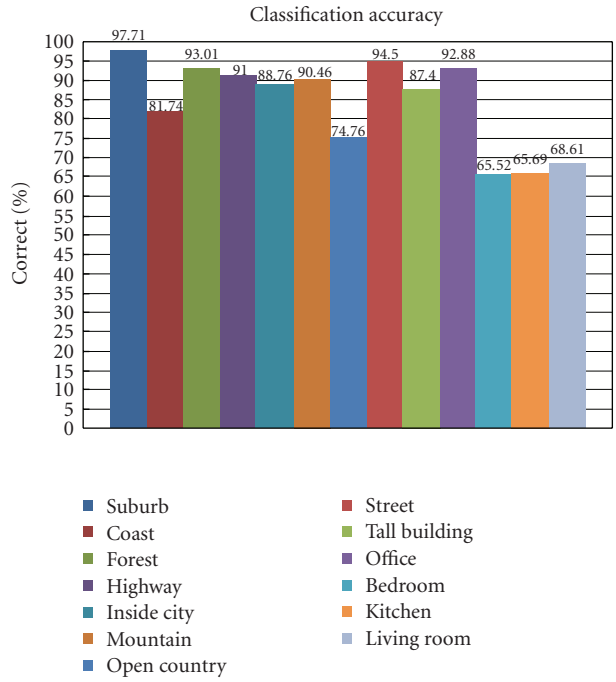

Figure 7: Classification accuracy on the thirteen basic categories used in [6] obtained considering the proposed representation and SVM.

Table 6: Rank statistics of the two best choices on the fifteen basic classes obtained considering the proposed representation and SVM.

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

Finally, the proposed representation coupled with SVM outperforms the results obtained in our previous work [31] where KNN was used together with the similarity measure defined in Section 3.2. In [31] the overall classification rate was 75.07% considering the ten basic classes (Accuracy is 14% less than the ones obtained using SVM on the same dataset.), 90.06% considering the superordinate level of description *In* versus *Out*, and 93.4% considering the superordinate level of description *Natural* versus *Artificial*. Confusion Matrix obtained using KNN are reported in Tables 7, 8, and 9. As shown by Table 10, the proposed similarity measure achieves better results with respect to other similarity measures.

In Figure 8 are reported some examples of images classified employing a K-nearest neighbors and the similarity measure described in Section 3.2. In particular the images to be classified are depicted in the first column, whereas the first three closest images used to establish the proper class of test image are reported in the remaining columns. The results are semantically consistent in terms of visual content (and category) to the related images to be classified.

5. Conclusion and Future Works

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 weighted bags of visual words representation is coupled with a discriminative method to perform classification. 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 complexity allowing at the same time to obtain effective

Figure 8: Examples of images classified with KNN and the similarity measure based on Bhattacharyya coefficient. The test images are on the left, and top three closest images used for classification are shown on the right.

Table 7: Confusion matrix obtained considering the proposed representation and KNN on the basic level of description of the scenes. The average classification rates for individual classes are listed along the diagonal.

	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	**63.34**	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	**49.06**	31.36	5.08	8.33	0.00	0.42
Open country	0.58	0.00	1.01	4.60	1.58	**76.68**	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 8: *Natural* versus *Artificial* results obtained considering the proposed representation and KNN classifier.

	Natural	Artificial
Natural	**92.88**	7.12
Artificial	5.98	**94.02**

Table 9: *In* versus *Out* obtained considering the proposed representation and KNN classifier.

	Out	In
Out	**91.63**	8.37
In	11.50	**88.50**

results. Future works should be devoted to perform a depth comparison between different kinds of features used to build the visual vocabulary (e.g., Textons versus SIFT) for scene classification. Moreover, since subregions characterized by different visual appearance but similar statistics of visual words may be confused in the proposed model, future works will be devoted in augmenting the model to capture the co-occurrences of visual words by means of correlograms taking

Table 10: Classification accuracy taking into account different similarity measures used by K-nearest neighbors algorithm. The similarity measure based on Bhattacharyya coefficient outperforms the other similarity measures in terms of classification accuracy.

Similarity measure	Accuracy
Bhattacharyya	*75.07*
χ^2	72.51
Absolute difference	71.30
Kullback-Leibler	71.14
Jeffrey	71.28
Euclidean	56.14

into account spatial constraints (like correlatons [46]) and computing the relationship between visual words directly on the feature domain. Although recent advances in the field, different challenges are still open in this research area; among others we highlight the following topics.

(i) Studies on how to produce powerful visual vocabularies to better discriminate between different classes are becoming more appealing [47, 48].

(ii) Models that exploit local and global information to better discriminate complex scenes environments (e.g., indoor scenes) are under consideration [49].

(iii) Very large datasets of images are already available [49–51] and there is the need to develop advanced techniques to scale with large dataset [52].

References

[1] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[2] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

[3] S. Battiato, G. M. Farinella, G. Giuffrida, C. Sismeiro, and G. Tribulato, "Using visual and text features for direct marketing on multimedia messaging services domain," *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 5–30, 2009.

[4] S. Battiato, G. M. Farinella, G. Giuffrida, C. Sismeiro, and G. Tribulato, "Exploiting visual and text features for direct marketing learning in time and space constrained domains," *Pattern Analysis and Applications*, vol. 13, no. 2, pp. 1–15, 2010.

[5] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *Proceedings of the European Conference on Computer Vision (ECCV '06)*, July 2006.

[6] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 524–531, June 2005.

[7] L. W. Renninger and J. Malik, "When is scene recognition just texture recognition?" *Vision Research*, vol. 44, pp. 2301–2311, 2004.

[8] P. Ladret and A. Guérin-Dugué, "Categorisation and retrieval of scene photographs from a JPEG compressed database," *Pattern Analysis and Applications*, vol. 4, no. 2-3, pp. 185–199, 2001.

[9] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 2169–2178, 2006.

[11] M. A. Tahir, K. V. de Sande, J. Uijlings, et al., "SurreyUVA SRKDA method, university of amsterdam and university of surrey at pascal voc 2008," in *Proceedings of the Visual Object Classes ChallengeWorkshop, in Conjunction with IEEE European Conference on Computer Vision*, 2008.

[12] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer, "Learning object representations for visual object class recognition," in *Proceedings of the ChallengeWorkshop in Conjunction with IEEE European Conference on Computer Vision (ICCV '07)*, 2007.

[13] K. E. A. Van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. In press.

[14] M. R. Boutell and J. Luo, "Beyond pixels: exploiting camera metadata for photo classification," *Pattern Recognition*, vol. 38, no. 6, pp. 935–946, 2005.

[15] J. Shawe-Taylor and N. Cristianini, *Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK, 2000.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA, 2006.

[17] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '03)*, vol. 2, pp. 1470–1477, 2003.

[18] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, no. 5802, pp. 91–97, 1981.

[19] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision*, vol. 62, no. 1-2, pp. 61–81, 2005.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[21] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions," *Bulletin of Calcutta Mathematical Society*, vol. 35, 1943.

[22] D. Marr and W. H. Freeman, Vision, 1982.

[23] I. Biederman, "Aspects and extension of a theory of human image understanding," in *Computational Processes in Human Vision: An Interdisciplinary Perspective*, Z. Pylyshyn, Ed., Ablex, Norwood, NJ, USA, 1998.

[24] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *Journal of the Optical Society of America. A*, vol. 7, no. 5, pp. 923–932, 1990.

[25] A. Oliva and A. Torralba, "Chapter 2 building the gist of a scene: the role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[26] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '03)*, vol. 1, pp. 273–280, Washington, DC, USA, 2003.

[27] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1226–1238, 2002.

[28] S. Battiato, S. Curti, M. La Cascia, M. Tortora, and E. Scordato, "Depth map generation by image classification," in *Three-Dimensional Image Capture and Applications VI*, vol. 5302 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2004.

[29] J. Vogel, A. Schwaninger, C. Wallraven, and H. H. Bülthoff, "Categorization of natural scenes: local versus global information and the role of color," *ACM Transactions on Applied Perception*, vol. 4, no. 3, 2007.

[30] G. M. Farinella, S. Battiato, G. Gallo, and R. Cipolla, "Natural versus artificial scene classification by ordering discrete fourier power spectra," in *Proceedings of 12th International Workshop on Structural and Syntactic Pattern Recognition, (SSPR '08), Satellite event of the 19th International Conference of Pattern Recognition (ICPR '08)*, Lecture Notes in Computer Science, pp. 137–146, 2008.

[31] S. Battiato, G. M. Farinella, G. Gallo, and D. Ravì, "Scene categorization using bag of textons on spatial hierarchy," in *Proceedings of the International Conference on Image Processing (ICIP '08)*, pp. 2536–2539, 2008.

[32] M. Gorkani and R. Picard, "Texture orientation for sorting photos "at a glance"," in *Proceedings of the IEEE Conference on Pattern Recognition (ICPR '94)*, vol. 1, pp. 459–464, October 1994.

[33] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *Proceedings of the IEEE International Workshop on Content-based Access of Image and Video Databases, in Conjunction with (ICCV '98)*, pp. 42–51, 1998.

[34] A. Bosch, X. Muñoz, and R. Martí, "Review: which is the best way to organize/classify images by content?" *Image and Vision Computing*, vol. 25, no. 6, pp. 778–791, 2007.

[35] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[36] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[37] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1458–1465, Washington, DC, USA, 2005.

[38] C. Harris and M. Stephens, "A combined corner and edge detection," in *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151, 1988.

[39] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[40] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka, "Visual categorization with bags of keypoints," in *Proceedings of the International Workshop on Statistical Learning in Computer Vision (ECCV '04)*, 2004.

[41] F. Porikli, "Integral histogram: a fast way to extract histograms in cartesian spaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 829–837, June 2005.

[42] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1800–1807, IEEE Computer Society, Beijing, China, October 2005.

[43] M. Johnson, *Semantic segmentation and image search*, Ph.D. thesis, University of Cambridge, Cambridge, UK, 2008.

[44] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.

[45] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," 2001.

[46] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 2033–2040, 2006.

[47] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[48] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," in *Neural Information Processing System (NIPS)*, B. Schőlkopf, J. Platt, and T. Hoffman, Eds., pp. 985–992, MIT Press, Cambridge, Mass, USA, 2006.

[49] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 413–420, Miami, Fla, USA, June 2009.

[50] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: a large data set for nonparametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[51] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep. 7694, California Institute of Technology, Pasadena, Calif, USA, 2007.

[52] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.