Available online at www.sciencedirect.com
ScienceDirect
Image and Vision Computing xxx (2006) xxx–xxx
www.elsevier.com/locate/imavis

A review: Which is the best way to organize/classify images by content?

Anna Bosch *, Xavier Muñoz, Robert Martí

Department of Electronics Informatics and Automatics, University of Girona, Campus Montilivi, Edifici P IV, Av. Lluis Santaló, s/n 17071 Girona, Spain

Received 19 December 2005; received in revised form 13 June 2006; accepted 12 July 2006

Abstract

Thousands of images are generated every day, which implies the necessity to classify, organise and access them using an easy, faster and efficient way. Scene classification, the classification of images into semantic categories (e.g. coast, mountains and streets), is a challenging and important problem nowadays. Many different approaches concerning scene classification have been proposed in the last few years. This article presents a detailed review of some of the most commonly used scene classification approaches. Furthermore, the surveyed techniques have been tested and their accuracy evaluated. Comparative results are shown and discussed giving the advantages and disadvantages of each methodology.

Keywords: Scene classification; Object recognition; Semantic concepts; Image segmentation

1. Introduction

Scene classification has the aim of labelling automatically an image among a set of semantic categories (e.g. coast, mountain and street). Fig. 1 shows several representative categories and images to illustrate this topic. It is an important task which helps to provide contextual information to guide other processes such as object recognition [1]. From an application viewpoint, scene classification is relevant in systems for organisation of personal and professional imaging collections, and has been widely explored in content based image retrieval systems [2–4]. Scene classification is valuable in image retrieval from databases because an understanding of the scene content can be used for efficient and effective database organisation and browsing. In addition, image filtering and enhancement operations may be adjusted depending on the scene type, so that the best rendering can be achieved.

This goal is not as ambitious as the general image understanding problem which tries to recognise every object in the image. Scenes can be often classified without having a full knowledge of every object. In some cases the use of low-level information, such as colour and texture, might be enough to classify some scenes. However in complex applications, although object recognition might be necessary, probably it is sufficient with a coarse recognition of not necessarily every object in the image. For instance, if a person sees trees at the top of an image and grass at the bottom, he can hypothesise that he is looking at a forest scene, even if he can not see every detail in the image [5] (Fig. 2).

Are image features enough to describe an scene or do we need to know which objects are present? The problem of scene modelling for classification using low-level features has been studied in image and video retrieval for several years [6]. Pioneering works used colour, texture and shape features directly from the image in combination with supervised learning methods to classify images into several semantic classes (indoor, outdoor, city, landscape, sunset, forest, ...). On the other hand, the modelling of scenes by a semantic intermediate representation was next proposed in order to reduce the gap between low-level and high-level image processing, and therefore to match the scene model with the perception we humans have (e.g. a street scene mainly contains road and buildings). The way to model an scene contains one of the main criteria

* Tel.: +34 972418891.
E-mail address: aboschr@eia.udg.es (A. Bosch).

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2006.07.015

Fig. 1. Examples of images used for the scene classification problem: coast, forest, mountain, highway, street and kitchen.

when identifying basic strategies which tackle the scene classification problem. Hence, the answers that authors have proposed to this modelling reveal these two major approaches: **low-level** and **semantic** modelling.

Furthermore, the question of whether feature information is sufficient is still open nowadays. Thorpe et el. [7] found that humans are able to categorise complex natural scenes containing animals or vehicles very quickly. Fei-Fei et al. [8] later showed that little or no attention is needed for such rapid natural scene categorisation. Both of theses studies posed a serious challenge to the currently accepted view that to understand the context of a complex scene, one needs first to recognise the objects and then in turn recognise the category of the scene [9]. Moreover, recent proposals have extended the meaning of semantic modelling to semantic concepts further than objects.

In this work, we will review the most recent and significant works in the literature on scene classification. Besides, we consider that the high number and diversity of recent proposals make necessary a finer classification than the classical two class modelling strategy. Hence, we have identified key approaches based on that criteria and we have classified the analysed works. Among the low-level methods, we distinguish between those that model the image as a single object, and those that partition the image in sub-blocks. Among the semantic methods, we distinguish three different approaches according to the meaning they give to the semantic of scenes, and hence which is the representation they build: techniques which describe the image by the objects and those that build the semantic representation from local information, and proposals which describe the image by semantic properties. Besides, we have implemented different algorithms in order to carry out a quantitative evaluation and a comparison of these approaches over a wide dataset.

The paper is structured as follows: firstly, we define and classify methods that use low-level features which will be referred to as global methods (Section 2), and then methods based on a semantic modelling (Section 3). We analyse in depth the review methods, what kind of features are employed, and the number of scene categories that the systems are able to recognise. Next, a quantitative evaluation of different approaches is shown in Section 4, along with the discussion of the results. A summary and conclusions from this work end this paper.

2. Low-level scene modelling

The problem of scene categorisation is often approached by computing low-level features (e.g. colour and texture), which are processed with a classifier engine for inferring high-level information about the image. These methods consider therefore that the type of scene can be directly described by the colour/texture properties of the image. For instance, a forest scene presents highly textured regions (trees), a mountain scene is described by an important amount of blue (sky) and white (snow), or the presence of straight horizontal and vertical edges denotes an urban scene.

A number of recent studies have presented approaches to classify indoor vs outdoor, or city vs landscape, using global cues (e.g. power spectrum, colour histogram information). Among them it is possible to distinguish two trends:

(1) ***Global:*** the scene is described by low-level features from the whole image.

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

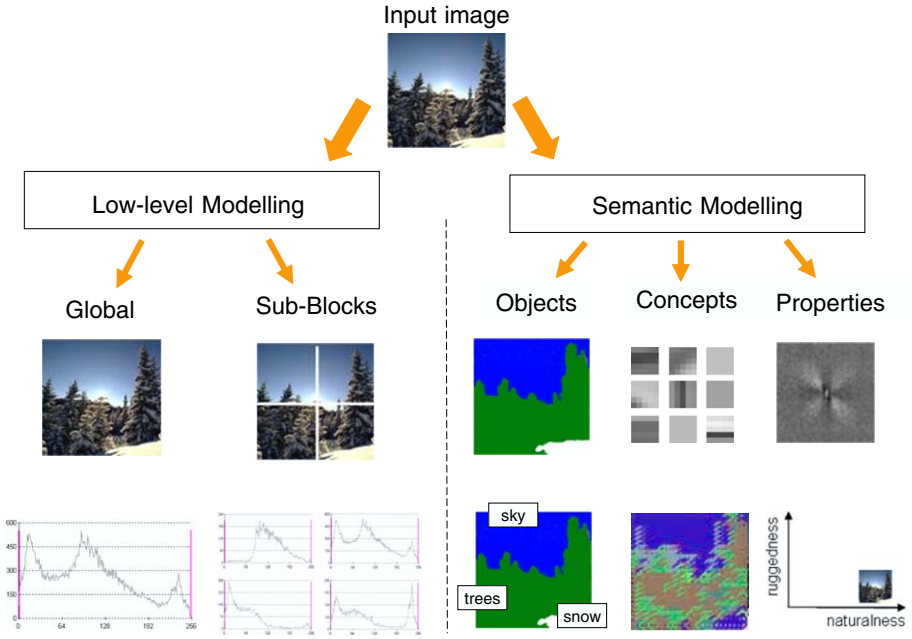

Fig. 2. Scene classification approaches. Low-level or semantic modelling, is the main property to distinguish basic strategies to tackle the proposed classification. Several approaches have been identified in both main strategies depending on how they achieve the final scene classification.

(2) ***Sub-blocks:*** the image is first partitioned into several blocks, and then features are extracted from each of those blocks.

In this section a review of the most recent and representative proposals of both global and sub-block approaches is presented.

2.1. Global

Vailaya et al. [10–12] consider the hierarchical classification of vacation images, and show that low-level features can successfully discriminate between many scenes types using a hierarchically structure. Using binary Bayesian classifiers, they attempt to capture high-level concepts from low-level image features under the constraint that the test image belongs to one of the classes. At the highest level, images are classified as indoor or outdoor; outdoor images are further classified as city or landscape; finally, a subset of landscape images is classified into sunset, forest, and mountain classes. Different qualitative measures, extracted from the whole image, are used at each level depending on the classification problem: indoor/outdoor (using spatial colour moments); city/landscape (edge direction coherence vectors), and so on. The classification problem is addressed by using Bayes decision theory. Each image is represented by a feature vector extracted from the image. The probabilistic models required for the Bayesian approach are estimated during a training step. Consider n training samples from a class w. A vector quantiser is used to extract q codebook vectors, v_j, from the n training samples. The class-conditional density of a feature vector y given the class w, i.e., $f_Y(y|w)$, is then approximated by a mixture of Gaussians (with identity covariance matrices), each centered at a codebook vector, resulting in:

$$f_Y(y|w) \propto \sum_{j=1}^{q} m_j \exp(-\|y - v_j\|^2/2) \quad (1)$$

where m_j is the proportion of training samples assigned to v_j. The Bayesian classifier is then defined using the maximum a posteriori (MAP) criterion as follows:

$$\hat{w} = \arg\max_{w\in\Omega}\{p(w|y)\} = \arg\max_{w\in\Omega}\{f_Y(y|w)p(w)\} \quad (2)$$

where Ω is the set of pattern classes and $p(w)$ represents the a priori class probability. The proposal reports an excellent performance at each level of the hierarchy over a set of 6931 images. However, it suffers a limitation inherent to hierarchical classifiers that is the cascading of errors. To classify a test image, for example a forest, into a category implies that we have to successfully classify the image at several stages (1) outdoor, (2) landscape, and (3) forest, with the probability of missing at each level. And obviously an initial mistake can not be solved at lower levels.

Also in [13] global features are used to produce a set of semantical labels with a certain belief for each image. They manually label each training image with a semantic label and train k classifiers (one for each semantic label) using support vector machines (SVM). Each test image is classified by the k classifiers and assigned a confidence score for the label that each classifier is attempting to predict. As a result, a k-nary label-vector consisting of k-class membership is generated for each image. This approach is specially useful for Content Based Image Retrieval (CBIR) and Relevance Feedback (RF) systems. Other authors have followed this global approach, although they have taken other aspects into account. For example, Shen et al. [14]

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

makes emphasis on the type of features that must be used. The authors argue that due to the complexity of visual content, a classification system can not be achieved by considering only a single type of feature such as colour, texture and shape alone and proposed Combined Multi-Visual Features. It produces a low-dimensional feature vector useful for an effective classification. Their method is tested on image classification using three different classifiers: SVM, K-Nearest Neighbours (K-NN) and Gaussian Mixture Models (GMM).

2.2. Sub-blocks

The scene can also be modelled by low-level features, but not from a single, whole image representation. Several proposals first split the image into a set of subregions, which are independently described by their low-level properties. These blocks are then classified, and finally the scene is categorised from the individual classification of each block.

The origin of this approach can be found in 1997, when Szummer and Picard [15] proposed to independently classify image subsections to obtain a final result using a majority voting classifier. The goal of this work was to classify images as indoor or outdoor. The image is first partitioned into 16 sub-blocks from which Ohta-space colour histograms and MSAR texture features are then extracted. K-NN classifiers are employed to classify each sub-block using the histogram intersection norm, which measures the amount of overlap between corresponding buckets in the two N-dimensional histograms X and Y and is defined as:

$$\mathrm{dist}(X,Y) = \sum_{i=1}^{N}(X(i) - \min(X(i), Y(i))) \quad (3)$$

Finally the whole image is classified using a majority voting scheme from the sub-block classification results. They obtain a 90.3% of performance, showing how high-level scene properties can be inferred from classification of low-level image features, specifically for the indoor/outdoor scene retrieval problem. They also demonstrated that performance is improved by computing features on sub-blocks, classifying these sub-blocks, and then combining theses results in a way reminiscent of stacking. Similar results were also obtained by Paek and Chang [16]. Moreover, they developed a framework to combine multiple probabilistic classifiers in a belief network. They trained classifiers for indoor/outdoor, and sky/no sky and vegetation/no vegetation as secondary cues for the indoor/outdoor problem. The classification results of each one are then feeded into a belief network to take the integrated decision.

The proposal of Serrano et al. [17] in 2004 shares this same philosophy, but using SVM for a reduction in feature dimensionality without compromising classification accuracy. Also colour and texture features are extracted from image sub-blocks and separately classified. Thus indoor/outdoor labels are obtained for different regions of the scene. The advantage of using SVM instead of K-NN classifier is that the sub-block beliefs can be combined numerically rather than by majority voting, which minimises the impact of sub-blocks with ambiguous labelling.

Even the good performance obtained by above proposals, one problem with the methods using image features for scene categorisation is that it is often difficult to generalise theses methods to additional image data beyond the training set. More importantly, they lack of an intermediate semantic image description that can be extremely valuable in determining the scene type. Hence, we draw our attention to systems that do attempt to find objects or other semantic concepts.

3. Intermediate semantic modelling

Scene content such as the presence of people, sky, grass, etc. may be used as cues for improving the classification performance obtained by low-level features alone [17], allowing to deal with the gap between low- and high-level features. Thus, an intermediate representation which models the content of the image is posteriorly used for scene classification. This intermediate representation is refered to as *semantic modelling*. Fig. 3 shows a graphical example. In this case, there are five semantic local concepts, and each "patch" of the image is assigned to one of them. Subsequently, images are classified as belonging to a certain scene according to their semantic concepts distributions. Note that in this Figure, all the images are road scenes and have similar semantic concepts distributions. In this case, local semantic concepts were obtained using probabi-

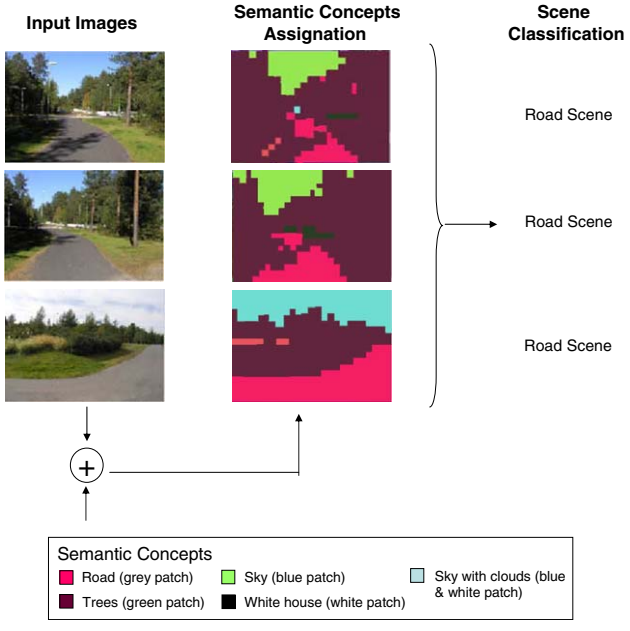

Fig. 3. Scene classification using semantic modelling: (i) each patch of the image is classified as a local semantic concept, (ii) each image is classified as an scene. Images are obtained from the Outex dataset [19].

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

listic Latent Semantic Analysis – pLSA (see Section 3.2.1) and we limited the images to road scenes to simplify the schema.

Nevertheless, the meaning of the semantic of the scene is not unique, and authors have proposed different semantic representations of images. A classical approach identifies the semantic as the set of objects that appear in the image (e.g. sky, grass, and mountains), and the scene is described by the occurrence of these *semantic objects*. Hence, these works imply an initial object detection step in the images. Some recent proposals try to avoid object segmentation and detection, and use more general intermediate representations. In this case, they first identify a dictionary of visual words or *local semantic concepts*, and further learn the visual words distribution for each scene category. Local semantic concepts define the semantic of the image from local information. These local concepts generally identify objects like *blue sky*, *gray sky*, *water with waves*, *mountain with snow*, or *mountain without snow* (this will be further analysed in Section 4.3). Furthermore, the last semantic definition can be clearly exemplified by the work of Oliva and Torralba [18] where the scene is described by local and global qualities related to the scene structure such as ruggedness, expansiveness, etc. We refer to these methods as the ones which use *Semantic Properties*. Following we summarise the three meanings of semantic modeling:

(1) ***Semantic Objects:*** objects of the image are detected to describe the scene. It mainly relies on an initial segmentation of the image into meaningful regions. Next, regions are labelled as known objects (semantic objects).
(2) ***Local Semantic Concepts:*** semantic of the image is represented by intermediate properties extracted from local descriptors around points.
(3) ***Semantic Properties:*** semantic of the image is described by a set of statistical properties/qualities of the image, such as naturalness, openness and roughness.

The intermediate semantic modelling makes more difficult the problem we are tackling because it habitually involves a local/region processing like (a not necessarily accurate) object recognition. On the other hand, it provides a potentially larger amount of information that must be exploited to achieve a higher performance on scene classification.

In this section, we review and classify different recent methods proposed in the literature which apply a semantic strategy.

3.1. Semantic objects

These methods are mainly based on first segmenting the image in order to deal with different regions. Subsequently local classifiers are used labelling the regions as belonging to an object (e.g. sky, people, cars, grass, etc.). Finally, using this local information, the global scene is classified. Different ways to carry out the scene classification using this strategy have been proposed recently.

Fan et al. [20] used concept sensitive salient objects as the dominant image components to achieve automatic image annotation at a content level. To detect the concept-sensitive salient objects, a set of detection functions is learned from the labelled image regions and each function is able to detect a specific type of these salient objects. Each detection function consists of three parts: (i) automatic image segmentation by using the mean shift technique, (ii) binary image region classification by using the SVM classifiers with an automatic scheme for searching the optimal model parameters and (iii) label-based aggregation of the connected similar image regions for salient object generation. To generate the semantic image concepts, the finite mixture models are used to approximate the class distributions of the relevant objects. After detecting the semantic salient objects they carry out the semantic image classification. The class distribution of these concept-sensitive salient objects $I_l = S_1, S_2, \ldots, S_n$ is then modeled as a finite mixture model $P(X, C_j|k, W_{cj}, \Theta_{cj})$. The test image I_l is finally classified into the best matching semantic scene concept C_j with the maximum posterior probability:

$$P(C_j|X, I_l, \Theta) = \frac{P(X, C_j|k, w_{cj}, \Theta_{cj})P(C_j)}{\Sigma_{j=1}^{N_c} P(X, C_j|k, W_{cj}, \Theta_{cj})P(C_j)} \quad (4)$$

where N_c is the number of classes (semantic scene concepts), w_{cj} is the set of the relative weights among the multivariate mixture components, X is the n-dimensional visual features that are used for representing the relevant concept-sensitive salient objects, k indicates the optimal number of multivariate mixture components, and $P(C_j)$ is the prior probability of the semantic image concept C_j in the database. $\Theta = k, w_{cj}, \Theta_{cj}, j = 1, \ldots, N_c$ is the set of mixture parameters and relative weights for the classifiers. An adaptative EM algorithm has been proposed to determine the optimal model structure and model parameters simultaneously. In addition, a large number of unlabelled samples are integrated with a limited number of labelled samples to achieve more effective classifier training and knowledge discovery.

Luo et al. [21] proposed an hybrid approach: low-level and semantic features are integrated into a general-purpose knowledge framework that employs a Bayesian Network (BN). BN are directed, acyclic graphs that encode the cause-effect and conditional independence relationships among variables in the probabilitic reasoning system. The directions of the links between the nodes (variables) represent causality in the sense that those links express the conditional probabilities of inferring the existence of one variable given the existence of the other variable. Each node can have many such directed inputs and outpus, each specifying its dependence relationship to the nodes from which the inputs originate (parents) and nodes where the outputs go (children). According to the Bayes rule, the pos-

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

terior probability can be expressed by the joint probability, which can be further expressed by the conditional and prior probabilities:

$$P(S|E) = \frac{P(S,E)}{P(E)} = \frac{P(E|S)P(S)}{P(E)} \quad (5)$$

where S denotes the semantic task and E denotes evidence. The efficacy of this framework is demonstrated via three applications involving semantic understanding of pictorial images: (i) detection of the main photographic subjects in an image [22], (ii) selecting the most appealing image in an event, and (iii) classifying images into indoor or outdoor scenes. This last application refers specifically to the problem of scene classification [23]. The performance is quantitatively evaluated using only low-level features (Ohta colour space histograms and MSAR texture features as in [15]), and incorporating semantic features (sky and grass objects). They demonstrate that the classification performance can be significantly improved when semantic features are employed in the classification process.

Aksoy et al. [24] also applied a Bayesian framework in a visual grammar. Scene representation is achieved by decomposing the image into prototype regions and modelling the interactions between these regions in terms of their spatial relationships. Initially an image segmentation is performed using a classical split-and-merge algorithm. Then, the technique automatically learns representative region groups which discriminate different scenes and builds visual grammar models. Similarly, in [25], after segmenting the image into regions, features are extracted and regions classified. Finally, based on this local classification the algorithm classifies the entire image. Their main contribution is that they found that the addition of eigenregions (the principal components of the intensity of the region) to the feature vector improves region classification results and furthermore the image classification rates. A similar approach was proposed by Mojsilovic et al. [26] where authors first segment the image using colour and texture information to find the semantic indicators (e.g. skin, sky, water, etc.). Then, these objects are used to identify the semantic categories (i.e. people, outdoor, landscapes, etc.).

Finally, we can also include in this approach the proposal of Vogel and Schiele [27,28], although in this case the segmentation is performed by a simple spatial grid layout which splits the image into regular subregions. The technique uses both colour and texture to perform landscape scene classification and retrieval based on a two-stage system. First, the image is partitioned into 10×10 subregions, and each one is classified using K-NN or SVM. An image is then represented by a so-called concept occurrence vector (COV), which measures the frequency of different objects in a particular image. The average COV over all members of a category defines the category prototype (P_c):

$$P_c = \frac{1}{N_c} \sum_{j=1}^{N_c} \mathrm{COV}(j) \quad (6)$$

where c refers to one of the scene categories and N_c to the number of images in that category. Given this image representation, a prototypical representation for each scene category can be learnt. Scene classification is carried out by using the prototypical representation itself or Multi-SVM approaches.

3.2. Local semantic concepts

In the last years we can find in the literature on scene classification, an increasing number of proposals which make use of local semantic concepts. Hence, an intermediary semantic level representation is introduced as a first step between image properties and scene classification in order to deal with the semantic gap between low-level features and high-level concepts. Nevertheless, all these proposals do not rely on an initial segmentation. Otherwise, the content of the scene is described by local descriptors, for example codewords [29–32] as shown in Fig. 4. These methods have in common that work over the *bag-of-words*, a technique used for the statistical text analysis.

3.2.1. Bag-of-words

The bag-of-words methodology was first proposed for text document analysis and further adapted for computer vision applications. The models are applied to images by using a *visual* analogue of a *word*, formed by vector quantising visual features (colour, texture, etc.) like region descriptors. Recent works have shown that local features represented by bags-of-words are suitable for scene classification showing impressive levels of performance [33–36].

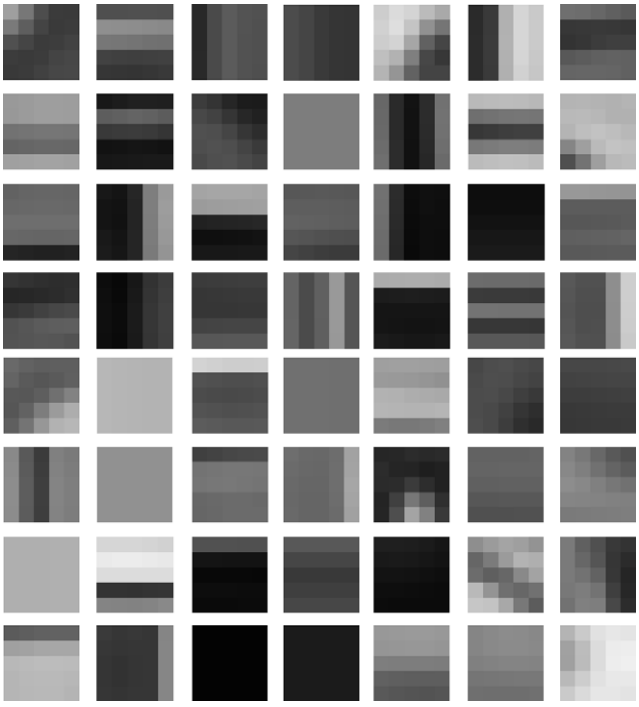

Fig. 4. Local descriptors represented by 5×5 patches at greyscale.

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

Constructing the bag-of-words from the images involves the following steps: (i) Automatically detect regions/points of interest, (ii) compute local descriptors over those regions/points, (iii) quantise the descriptors into words to form the visual vocabulary, (iv) find the occurrences in the image of each specific word in the vocabulary in order to build the bag-of-words (histogram of words). Fig. 5 schematically describes the four steps involved in the definition of the bag-of-words model.

Some Bayesian text models, such as probabilistic Latent Semantic Analysis (pLSA) [37] and Latent Dirichlet Analysis (LDA) [38,39] have been adapted and used to model scene categories. In text analysis they are used to discover topics in a document using the bag-of-words document representation. Here we have *images* as *documents* and we discover *topics as object categories* (e.g. *grass*, *houses*, *blue sky*, *gray sky*), so that an image containing instances of several objects is modelled as a mixture of topics. This topics distribution over the images is used to classify an image as belonging to a certain scene (e.g. if an image contains *water with waves*, *sky with clouds* and *sand* will be classified as a *coast* scene).

Suppose that we have a collection of images $D = d_1,\ldots,d_N$ with words from a visual vocabulary $W = w_1,\ldots,w_V$. One may summarize the data in a $V \times N$ co-occurrence table of counts $N_{ij} = n(w_i, d_j)$, where $n(w_i, d_j)$ denotes how often the word w_i occurred in an image d_j. In pLSA there is also a latent variable model for co-occurrence data which associates an unobserved class variable $z \in Z = z_1,\ldots,z_Z$ with each observation. A joint probability model $P(w,d)$ over $V \times N$ is defined by the mixture:

$$P(w|d) = \sum_{z\in Z} P(w|z)P(z|d) \tag{7}$$

where $P(w|z)$ are the topic specific distributions and, each image is modelled as a mixture of topics, $P(z|d)$. The pLSA model is shown in Fig. 6a.

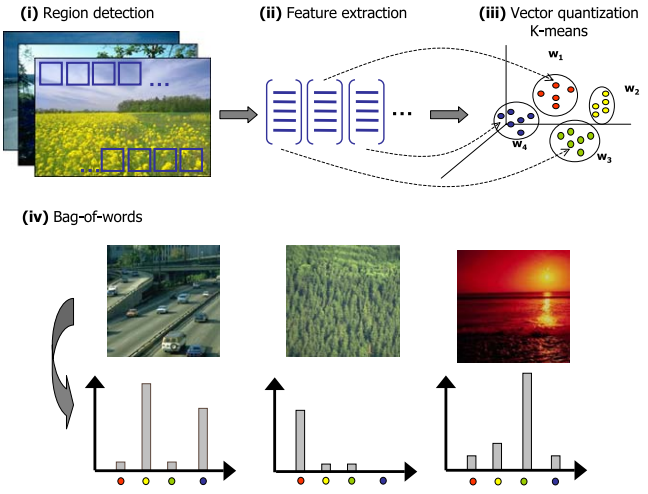

Fig. 5. Four steps to compute the bag-of-words when working with images. (i–iii) obtain the visual vocabulary by vector quantizing the feature vectors, and (iv) compute the image histograms – bag-of-words – for images according the obtained vocabulary.

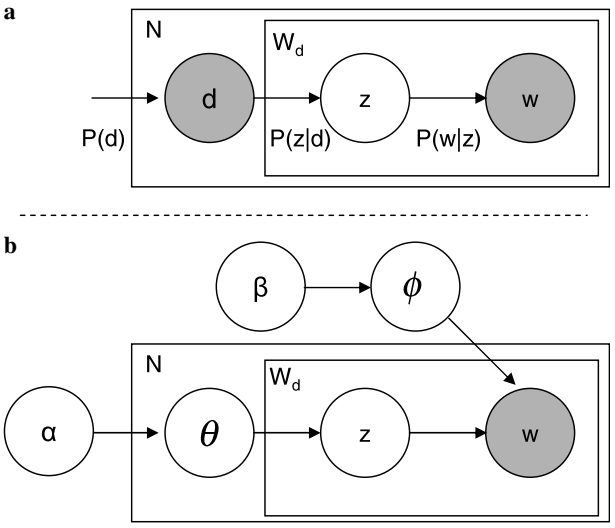

Fig. 6. (a) pLSA graphical model. Filled circles indicate observed random variables and the unfilled are unobserved, (b) LDA graphical model.

In contrast to pLSA, LDA treats the multinomial weights $P(z|d)$ over topics as latent random variables. The pLSA model is extended by sampling those weights from a Dirichlet distribution, the conjugate prior to the multinomial distribution [40]. This extension allows the model to assign probabilities to data outside the training corpus and uses fewer parameters, thus reducing overfitting. The LDA model is shown in Fig. 6b, where W_d is the number of words in document d. The goal is to maximize the following likelihood:

$$P(w|\phi,\alpha,\beta) = \int \sum_z P(w|z,\phi)P(z|\theta)P(\theta|\alpha)P(\phi|\beta)\mathrm{d}\theta \tag{8}$$

where θ and ϕ are multinomial parameters over the topics and words respectively and $P(\theta|\alpha)$ and $P(\phi|\beta)$ are Dirichlet distributions parameterized by the hyper parameters α and β.

Bosch et al. [33] provided an approach which uses bag-of-words to model visual scenes in image collections, based on local invariant features and pLSA. They successfully classified up to 13 categories out performing the state of the art on three known datasets (they report a 73.4% of correct classification). Quelhas et al. [35] used a similar approach presenting differences in terms of: (i) the number of scenes that the try to classify (3 in [35] and up to 13 in [33]), and (ii) how the features are used: In [35] SIFT descriptors are computed around an interest point – sparse descriptors, while in [33] SIFT features are computed on a regular grid and using concentric patches around each point to allow scale variance – dense descriptors. Moreover, it has been demonstrated that when working with scene classification, and concretely with natural images such as *coast* or *open country*, dense descriptors outperform the sparse ones (see also [34]).

Fei-Fei and Perona [34] independently proposed two variations of LDA firstly proposed by Blei et al. [38,39] which was designed to represent and learn document mod-

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

els. In this framework, local regions are first clustered into different intermediate themes (local semantic concepts), and then into categories. Probability distributions of the local regions as well as the intermediate themes are both learnt in an automatic way, bypassing any human annotation. No supervision is needed apart from a single category label to the training image. Performances are shown in a dataset consisting of 13 categories.

Recently, Perronnin et al. [41] defined a universal vocabulary, which describes the content of all the considered scenes, and class visual vocabularies which are obtained through the adaptation of the universal vocabulary using class-specific data. While previous approaches characterise an image with a single histogram, here an image is represented by a set of histograms, one per class. Each histogram describes whether an image is more suitably modelled by the universal vocabulary or the corresponding adapted vocabulary. They represent a vocabulary of visual words by means of a GMM where $\lambda = w_i,\ \mu_i,\ \Sigma_i,\ i = 1,\ldots,N$. λ denotes the set of parameters of a GMM, w_i, μ_i and Σ_i denote respectively the weight, mean vector and covariance matrix of Gaussian i and N denotes the number of Gaussians. Each Gaussian represents a word of the visual vocabulary. The Universal vocabulary is trained using maximum likelihood estimation (MLE) and the class vocabularyies are adapted using the maximum a posteriori (MAP) criterion. They successfully test the method classifying scene images from three different datasets consisting on scenes like sunrise/sunset, underwater, waterfalls, buildings, etc.

3.2.2. Bag-of-words with context

Habitual bag-of-words techniques, as the described above, do not take the spatial information into account. However, in complex natural images, scene classification systems can be further improved by using contextual knowledge like common spatial relationships between neighbouring local objects [42] or the absolute position of objects in certain scenes. While the above methods have shown to be effective, their neglect of spatial structure ignores valuable information which could be useful to achieve better results for scene classification.

In [36] they proposed a method for recognizing scene categories based on approximate global geometric correspondence. The technique works by partitioning the image into increasingly finer sub-regions and computing histograms of local features found inside each sub-region. The resulting spatial-pyramid is a simple computationally efficient extension of an orderless bag-of-words image representation. The scene classification is performed using a pyramid matching approach:

$$K_L(X,Y) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l \tag{9}$$

where X and Y are the two sets of vectors in a d-dimensional feature space, L is the grid resolution, such the grid at level l has 2^l cells along each dimension and I^l is the histogram intersection function at level l. A pyramid matching works by placing a sequence of increasingly coarser grids over the feature space and taking a weighted sum of the number of matches that occur at each resolution level. Multi-class classification is done with SVM. This method achieves high accuracy (a total of 81.4%) on a large database of 15 natural scene categories.

In [43], Fergus et al. develop two new models, ABS-pLSA and FSI-pLSA, which extend pLSA to include absolute position and spatial information in a translation and scale invariant manner respectively. Although this method is used for object classification, it could be easily adapted for scene classification tasks.

3.3. Semantic properties

Finally, the last group of works that make use of a semantic description as intermediate layer, has exploited the statistical properties of the scene. It breaks with a common trend of the above approaches since the semantic is here related to global configurations and scene structure, instead of local objects or regions. Consequently, neither segmentation nor the processing of local regions or objects is required. Therefore the image is described by visual properties, which are shared by images of a same category.

Oliva and Torralba [18,44,45] proposed a computational model for the recognition of real world scenes (four natural scenes and four man-made scenes) that bypasses the segmentation and the processing of individual objects or regions. The procedure is based on a very low dimensional representation of the scene, that they refer to as the Spatial Envelope. It consists of five perceptual qualities: naturalness (vs man-made), openness (presence of a horizon line), roughness (fractal complexity), expansion (perspective in man-made scenes), and ruggedness (deviation from the horizon in natural scenes). However, the contribution of each feature cannot be understood as they stand, and more importantly, they are not directly meaningful to human observers. Each feature corresponds to a dimension in the spatial envelope space, and together represent the dominant spatial structure of a scene. Then, they show that these dimensions may be reliably estimated using spectral and coarsely localised information. The model generates a multidimensional space in which scenes sharing membership in semantic categories are projected close together. Therein it is possible to assign a specific interpretation to each dimension: along the openness dimension, the image refers to an open or a closed environment, etc.

The estimation of each attribute s from the *global spectral features* v of a scene picture can be written as:

$$\hat{s} = v^T d = \sum^{N_G} v_i d_i = \int\int A(f_x,f_y)^2 DST(f_x,f_y)\mathrm{d}f_x\mathrm{d}f_y \tag{10}$$

with

$$\mathrm{DST}(f_x,f_y) = \sum^{N_G} d_i \Psi_i(f_x,f_y) \tag{11}$$

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

Eq. 10 shows that the spatial envelope property s is estimated by a dot product between the amplitude spectrum of the image and a template DST(f_x,f_y). The DST (discriminant spectral template) is a function that describes how each spectral component contributes to a spatial envelope property. The DST is parametrized by the column vector $d = d_i$ which is determined during a learning stage. A similar estimation can be performed when using the *spectogram features* w:

$$\hat{s} = w^T d = \sum^{N_L} w_i d_i = \sum_x \sum_y \int \int A(x,y,f_x,f_y)^2 \mathrm{WDST}(x,y,f_x,f_y)\mathrm{d}f_x\mathrm{d}f_y \quad (12)$$

with

$$\mathrm{WDST}(x,y,f_x,f_y) = \sum^{N_L} d_i \Psi_i(x,y,f_x,f_y) \quad (13)$$

The WDST (windowed discriminant spectral template) describes how the spectral components at different spatial locations contribute to a spatial envelope property. The performance of the spatial envelope model shows that specific information about object shape or identity is not a requirement for scene categorisation and that modelling a holistic representation of the scene informs about its probable semantic category. N_G and N_L are the number of functions used for the approximations and determine the dimensionality of each representation and $\Psi(f_x, f_y)$ are the KL basis of the energy spectrum.

4. Evaluation

A robust and objective methodology for the evaluation of existing approaches to scene classification is needed in order to discern the best method for an specific application field. Thus, albeit necessary, is not a trivial task due to the heterogenous data and classification implementations, and has often been misregarded in the existing literature on that specific topic. Proposals differ on objectives they try to satisfy (e.g. number and kind of scenes to classify), and the image data over they work with (specially constrained in some cases). Furthermore, test details as how the images were split into training and test sets are often not specified in published works. Hence, unless a given system is implemented and tested for specific image data, it is very difficult to evaluate from the published works how well it would work for that data.

It is our aim here, to provide and evaluation of although not all existing methodologies, the must representative works derived from each criteria reviewed so far. We designed and implemented three algorithms representative of the main approaches identified in this work. We then tested them over the same dataset used by Vogel and Schiele [27], which allowed us to compare their performance to the existing results published. We compared the results on scene classification obtained by four different methods mentioned above: (i) low-level image representation (LLI), (ii) low-level block representation (LLB), (iii) image segmentation by classifying present objects (IS) and (iv) bag-of-words model using pLSA (BOW). The Vogel and Schiele [28] dataset used includes 700 natural scenes from the Corel Database consisting of six categories: 144 coasts, 103 forests, 179 mountains, 131 open country, 111 river and 34 sky/clouds. The size of the images is 720×480 (landscape format) or 480×720 (portrait format). Every scene category is characterised by a high degree of diversity and presents potential ambiguities since it depends strongly on the subjective perception of the viewer. For example in Fig. 7a, the three *river* scenes could be also labelled as *forest* for someone, yet there is also a *forest* in these images. Moreover, we also evaluated the computational cost of the methods.

4.1. Features and methodology

We used 600 randomly selected training images and the rest for testing as in [27]. Features used are a concatenation of an 84 HSI histogram (with 36 bins for H, 32 bins for S, and 16 bins for V), 24 features of the gray-level co-occurrence matrices (32 gray levels): contrast, energy, entropy, homogeneity, inverse difference moment, and correlation, for the displacements $\overrightarrow{1,0}$, $\overrightarrow{1,1}$, $\overrightarrow{0,1}$, and $\overrightarrow{-1,1}$, and a 72-bin edge direction histogram. The final feature vector is then 180-dimensional. Moreover, we have evaluated optimum parameters values for each technique. Here, only the best results obtained with theses parameter values are shown. The methodologies for each strategy are the following:

(1) ***LLI:*** The algorithm computes global features for each training image, then each image is represented by a 180-dimensional vector. A test image is classified using K-NN (with $K = 10$).
(2) ***LLB:*** The algorithm extracts vector features for each block in the training image following the strategy proposed by Szummer and Picard [15]. We divided the image into 2×2 and 4×4 blocks. Each block from the test image is classified using K-NN (with $K = 10$) and then combining these results we classify the image by a majority voting.
(3) ***IS:*** This is the method implemented in [27]. They first classify each image patch (10×10 grid) providing from a certain object and using the object distribution the image classification is carried out. Authors in [27] worked with the same dataset and features as the used here in for evaluation. Thus we have used their published performance to compare it to other approaches in this paper.
(4) ***BOW:*** A 5×5 square neighbourhood around a pixel is used to compute the feature vector. The patches are spaced by 3 pixels on a regular grid. In this case we have lots of feature vectors and we quantize them using the k-means algorithm to form the visual

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

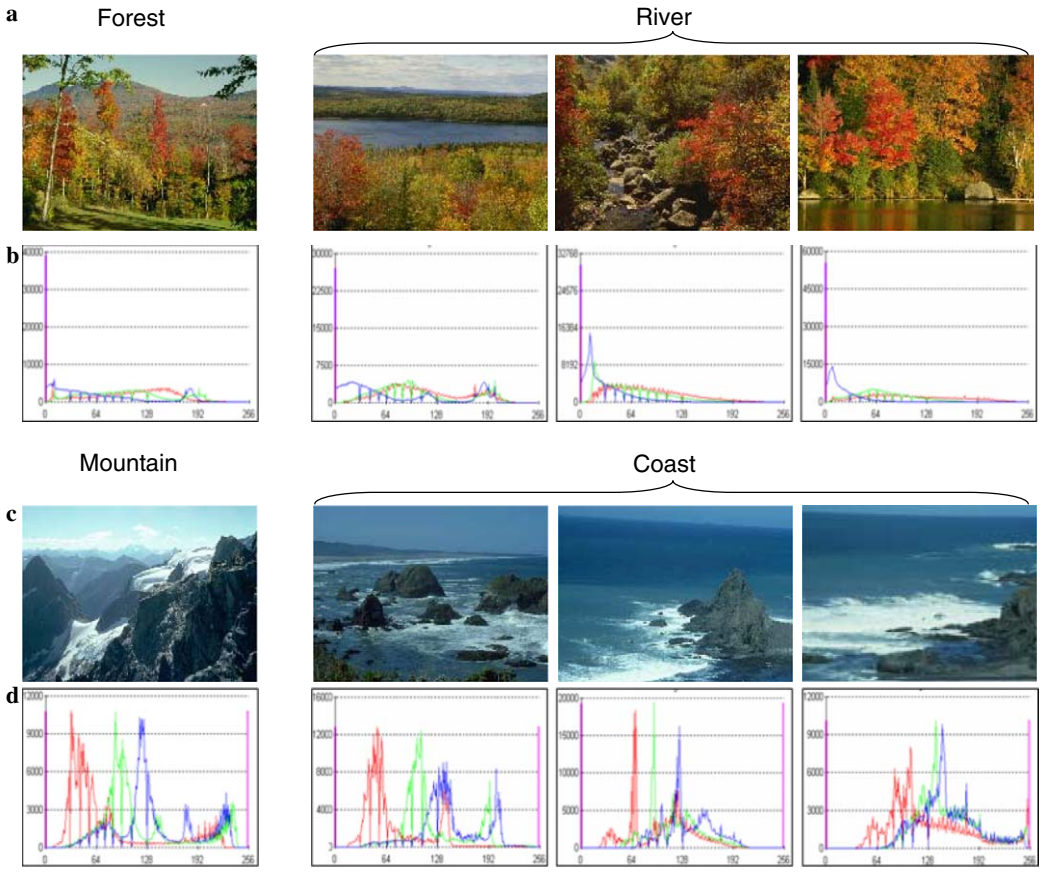

Fig. 7. Some typical scenes confused when using global method LLI for classification. (a) river images confused as forest, (c) coast images confused as mountain. (b and d) HSI histograms from the above images.

vocabulary. Then we classify each image using pLSA. We used $K = 700$ for the k-means algorithm (vector quantization) and $Z = 25$ when running pLSA.

The classification task is to assign each test image to one of the six categories. The performance is measured using a confusion table, and overall performance rates are measured by the average value of the diagonal entries of the confusion table.

4.2. Classification results

Classification results are shown in Table 1. As it is clearly stated, worse results were obtained by low-level approaches. A 53.25% of correct classification was achieved by the LLI algorithm, while a poor 49.12% was reached by the LLB algorithm. A low-level strategy, which considers the scene as an individual object, is normally used to classify only a small number of scene categories (indoor versus outdoor, city versus landscape, etc.). The six categories considered in our experiments are too complex to be distinguished by low-level scene properties.

Table 1
Performance of the compared approaches over a same dataset. LLI, LLB, BOW have been implemented by ourselves, while IS performance is the score published in [27]

LLI (%)	LLB (%)	IS (%)	Bow (%)
53.25	49.12	74.10	76.92

If we look at Fig. 7a, the ground truth of left image is *forest* while the ground truth of the other ones is *river*. However their colour and texture distribution is very similar (Fig. 7b shows the HSI histograms) and low-level method LLI fails when it tries to classify *river* scenes, classifying them as *forest* scenes. Something similar happens with Fig. 7c, where *coast* images are confused as *mountain*.

In fact, the set of images and categories used by most of the authors are often constrained. As an example, the categories used by Vaiyala [11] were chosen specifically to be nicely separable. The same author recognised in [11]: "we thus restricted classification of landscape images into three classes that could be more unambiguously distinguished, namely sunset, forest, and mountain classes. Sunset scenes can be characterised by saturated colours (red, orange or yellow), forest scenes have predominately green colour distribution due to the presence of dense trees and foliage, and mountain scenes can be characterised by long distance shots of mountains".

In contrast, when using local semantic concepts, or object segmentation, we can deal with objects (or concepts) in the images, and classify them in an easy way, according

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

to their distribution. Semantic techniques which make use of an intermediate representation achieved best the results, scores from 74% to almost 77% have been obtained (see Table 1). They can deal with the parts of the image that correspond to *trees*, and the ones that correspond to the *river*, and it will allow to distinguish between *forest* and *river* scenes (see Fig. 8). Note that we have not provided a comparison for the semantic properties-based methodologies, however it has been demonstrated in [33] that bag-of-words methods perform better than the representative proposal of Oliva and Torralba [18].

Computational Cost. Low-level strategies have two clear advantages: their simplicity and their low computational cost. Over the set of 600 training images, 3 min are needed to construct the classifier when using LLI, and 6 min and 40 s when LLB. This computational cost is much higher when using BOW. This is because we use more information from the images. We need a preprocessing step to construct the visual vocabulary which is a bit expensive: around 4 h extracting 6400 descriptors per image and running k-means with 700 clusters. The step for fitting pLSA takes 10 min. However, the costs to classify a test image are comparable: 2 and 7 s for LLI and LLB respectively and 15 s for BOW. Authors in [27] did not give the computational cost of their algorithm.

All above experiments have been done on a 1.7 GHz Computer and Matlab implementation.

4.3. Discussion

Although low-level strategies present a lowest computational cost, they have a poor performance, because they are unable to distinguish between complex scenes. Hierarchic schemes have been proposed to overcome this drawback, however our results seem to corroborate the inappropriate of these methods when the number of categories is increased.

On the other hand, the best results have been obtained when using Local Semantic Concepts with the bag-of-words and pLSA method (76.92%). Besides, this approach has a nice, very relevant property; local semantic approaches are also the ones which require less user intervention to learn "intermediate' representations: they directly learn from the data by an unsupervised (e.g. [35,33]) or semi-supervised (e.g. [34]) way. Contrarily, a main requirement of the other semantic modelling approaches is the manual annotation of these properties. In Oliva and Torralba work [18], human subjects are instructed to rank each of the hundreds of training scenes into 6 different properties. In [27], human subjects are asked to classify near 60,000 local patches from the training images into nine different "semantic concepts". Both cases involves tens of hours of manual labelling. Hence, a drawback of these strategies is their preprocessing cost, although this step could be done off-line.

Focusing on the semantic approaches, the main drawback when using segmentation techniques (IS) respect to the ones which use local semantic concepts (BOW), is probably the accuracy of the segmentation method. If objects are not well segmented all the posterior classification stages will probably fail. In contrast, when using local semantic concepts the segmentation process is omitted and the image is classified looking at the local patches. Furthermore, another important feature, to understand the best results obtained by BOW technique, is probably its freedom to choose appropriate concepts ($Z = 25$) for a dataset. The system organises them in his own way in order to have different object representa-

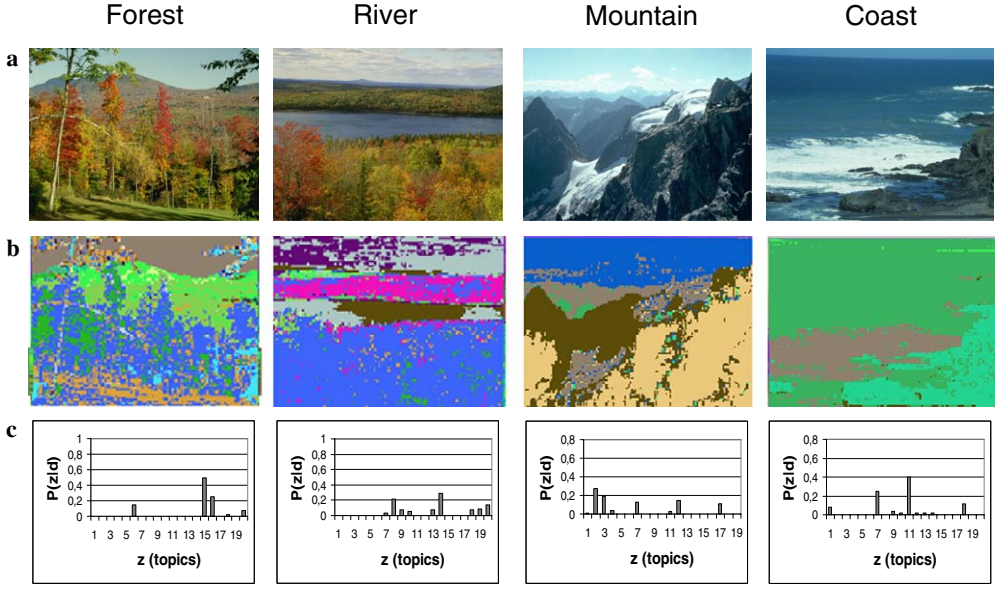

Fig. 8. Some scenes confused when using low-level methods and well classified when using local semantic concepts. (a) original image, (b) "topics" assigned by pLSA to each patch, (c) topic distribution – $P(z|d)$ – for each image. Note that previous confused images (see text) have different distributions according to its kind of scene, which allows to obtain a better classification rate.

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

Table 2
Summary of the analysed scene classification systems

	Author	Objects	Scenes	Features	#scenes
Low-level strategies					
Global	Vaialya et al. [10,12]	–	Bayesian classifiers	LUV and HSV color space (spatial moments, histograms, coherence vectors); MSAR; edge directions histograms, coherence vectors	5: indoor, city, sunset, forest and mountain
	Chang et al. [13]	–	SVM; Bayes point machine	Color, texture	15: architecture, bears, clouds, elephants, fabric, fireworks, flowers, food, landscape, object images, people, texture, tigers, tools, and waves
	Shen et al. [14]	–	SVM; K-NN; GMM	Color; texture; shape	10: natural scenery, architecture, plants, animals, rocks, flags, buses, food, human faces and roses
Sub-blocks	Szummer and Piccard [15]	–	K-NN; 3-layer NN; mixture of Experts classifier	Ohta color space; MSAR; shift-invariant DCT	2: indoor and outdoor
	Serrano et al. [17]	–	SVM	LST color space; wavelet texture	2: indoor and outdoor
Semantic strategies					
Objects	Fan et al. [20]	SVM	Bayesian framework	coverage ratio, region center, region rectangular box, Tamura texture, wavelet texture/color LUV,	6: mountain, view, beach, garden, saliling, skiing and desert
	Luo et al. [21,23]	K-NN	Bayesian Network	Ohta Color space, MSAR	2: indoor, outdoor
	Fredembach et al. [25]	Multivariate gaussin analysis based on the maximum a posteriori rule	Probabilistic method	RGB, Lab, co-occurrence matrix, amplitude spectrum of the fourier transform	3: vegetation, sky and skin
	Mojsilovic et al. [26]	Naive Bayes classifier	Bayesian framework	region spatial relationships	Portraits, people, outdoor, crowes, city, indoor, lanscapes, etc.
	Vogel and Schiele [27,28]	K-NN; SVM	M-SVM; SSD between category prototypes	Color (HSV,RGB) and edge histograms; co-occurrence matrix	6: sky, coast, mountains, field, river and forest
Concepts	Fei-Fei et al. [34]	Implicit with the method	Bag-of-words and LDA extension	Dense SIFT and gray level descriptors on a regular grid	13: forest, coast, mountain, open country, street, inside city, tall buildings, high way, bedroom, suburb, living room, kitchen and office
	Quelhas et al. [35]	Implicit with the method	bag-of-words and pLSA	sparse SIFT around interest points	3: indoor, city and landscape
	Bosch et al. [33]	Implicit with the method	Bag-of-words and pLSA	Dense SIFT on a regular grid – concentric patches allow scale variation	Up to 13: the same as in [34]
	Perronin et al. [41]	Implicit with the method	Bag-of-words and GMM	dense SIFT on a regular grid	Up to 10: Africa, beach, buildings, buses, dinousaurs, elephants, flowers, horses, mountains and food
	Lazebnik et al. [36]	Implicit with the method	Bag-of-words and pyramid kernels	Dense SIFT on a regular grid	15: the same as in [34] plus industrial and store
Propert.	Oliva and Torralba [18,44,45]	Implicit with the method	K-NN	Spatial envelope (DST WDST)	4 man-made: street, high way, tall buildings and inside city; 4 natural: forest, coast, mountain and open country

tion for the different scenes. For example, it distinguishes the sky as three different objects: *blue sky*, *grey sky*, and *sky with clouds*. It gives us additional information to classify scenes because image representation is more discriminative. In contrast, when using the semantic object-based method we classify the regions among nine objects, and most of these are in all the six categories (e.g. *sky* object), and makes the scene representation more ambiguous. We believe this freedom to select the most adequate concepts for each dataset is probably responsible of the superior performance of the scene classifier [33,46].

5. Summary and conclusions

For the purpose of providing an overview of the presented systems, Table 2 summarises the most relevant and recent approaches to scene classification. The first column identifies the different systems by giving authors names with referred papers. Next two columns refer to the strategy used in order to carry out the semantic concepts classification and the posterior scene classification. Finally, the last two columns summarise the features used and the number of scene categories classified in each paper.

We implemented and evaluated two low-level strategies as well as two approaches that use intermediate representations. We demonstrated that a better classification is achieved when a semantic representation is used in order to deal with the gap between low- and high-level. Low-level strategies are useful when a small number of categories have to be recognised, and also when the categories are easily separable. However, as the number and ambiguity of the categories increase it is clear that approaches using intermediate semantic concepts are more appropriate.

Latest trends are using the bag-of-words representations jointly with different techniques firstly proposed in the text document retrieval literature [47]. We also demonstrated that this method is the one which obtains the best classification results in our experiments. This approach provides a categorisation of individual features, and moreover in [48] it is shown that pLSA is also useful to model the object in the scene providing its *segmentation*. Besides, firstly proposed text models have been very recently extended to include contextual information in order to improve the scene classification. All these techniques are also very interesting due to the fact that they can discover the topic distribution even perform all the scene classification in an unsupervised way. In that sense, avoiding the tedious and time-consuming task of hand annotation and also the fact that expert-defined labels are somewhat arbitrary and possibly sub-optimal.

Acknowledgements

We thank Julia Vogel, for providing their image data and corresponding ground-truth. This work was partially funded by research grant BR03/01 from the University of Girona.

References

[1] A. Torralba, Contextual priming for object detection, International Journal of Computer Vision 53 (2) (2003) 169–191.
[2] F. Jing, M. Li, L. Zhang, H. Zhang, B. Zhang, Learning in region-based image retrieval, in: International Conference on Image and Video Retrieval, Urbana-Champaign, Illinois, 2003, pp. 206–215.
[3] W.H. Adams, G. Iyengar, C. Lin, M.R. Naphade, C. Neti, H.J. Nock, J.R. Smith, Semantic indexing of multimedia content using visual, audio, and text cues, Journal on Applied Signal Processing 2 (2003) 1–16.
[4] J. Wang, J. Li, G. Wiederhold, Simplicity: semantics-sensitive integrated matching for picture libraries, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (9) (2001) 947–963.
[5] M. Boutell, C. Brown, J. Luo, Review of the state of the art in semantic scene classification, Tech. rep., The University of Rochester, 2001.
[6] A.W. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.
[7] S. Thorpe, D. Fize, C. Marlot, Speed of processing in the human visual system, Nature 381 (1996) 520–522.
[8] L. Fei-Fei, R. VanRullen, C. Koch, P. Perona, Rapid natural scene categorization in the near absence of attention, Proceedings of the National Academy of Sciences of the United States of America 99 (14) (2002) 9596–9601.
[9] A. Treisman, G. Gelade, A feature-integration theory of atention, Cognitive Psychology 12 (1980) 97–136.
[10] A. Vailaya, A. Jain, H. Zhang, On image classification: city vs. landscapes, Pattern Recognition 31 (12) (1998) 1921–1935.
[11] A. Vailaya, M. Figueiredo, A. Jain, H. Zhang, Content-based hierarchical classification of vacation images in: IEEE International Conference on Multimedia Computing and Systems, vol. 1, Florence, Italy, 1999, pp. 518–523.
[12] A. Vailaya, A. Figueiredo, A. Jain, H. Zhang, Image classification for content-based indexing, IEEE Transactions on Image Processing 10 (2001) 117–129.
[13] E. Chang, K. Goh, G. Sychay, G. Wu, Cbsa: Content-based soft annotation for multimodal image retrieval using bayes point machines, IEEE Transactions on Circuits and Systems for Video Technology Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description 13 (1) (2003) 26–38.
[14] J. Shen, J. Shepherd, A.H.H. Ngu, Semantic-sensitive classification for large image libraries, in: International Multimedia Modelling Conference, Melbourne, Australia, 2005, pp. 340–345.
[15] M. Szummer, R.W. Picard, Indoor–outdoor image classification, in: IEEE International Workshop on Content-based Access of Image and Video Databases, in conjuntion with ICCV'98, Bombay, India, 1998, pp. 42–50.
[16] S. Paek, S.-F. Chang, A knowledge engineering approach for image classification based on probabilistic reasoning systems, in: IEEE International Conference on Multimedia and Expo, vol. II, New York, 2000, pp. 1133–1136.
[17] N. Serrano, A. Savakis, J. Luo, Improved scene classification using efficient low-level features and semantic cues, Pattern Recognition 37 (2004) 1773–1784.
[18] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, International Journal of Computer Vision 42 (3) (2001) 145–175.
[19] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, S. Huovinen, Outex – new framework for empirical evaluation of texture analysis algorithms in: IARP International Conference on Pattern Recognition, vol. 1, Québec City, 2002, pp. 701–706.
[20] J. Fan, Y. Gao, H. Luo, G. Xu, Statistical modeling and conceptualization of natural images, Pattern Recognition 38 (2005) 865–885.

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015

[21] J. Luo, A.E. Savakis, A. Singhal, A bayesian network-based framework for semantic image understanding, Pattern Recognition 38 (2005) 919–934.
[22] J. Luo, A. Singhal, S.P. Etz, R.T. Gray, A computational approach to determination of main subject regions in photographic images, Image and Vision Computing 22 (2004) 227–241.
[23] J. Luo, A. Savakis, Indoor vs outdoor classification of consumer photographs using low-level and semantic features, in: IEEE International Conference on Image Processing, vol. 2, Thessaloniki, Greece, 2001, pp. 745–748.
[24] S. Aksoy, K. Koperski, C. Tusk, G. Marchisio, J.C. Tilton, Learning Bayesian classifiers for scene classification with a visual grammar, IEEE Transactions on Geoscience and Remote Sensing 43 (3) (2005) 581–589.
[25] C. Fredembach, M. Schröder, S. Süsstrunk, Eigenregions for image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (12) (2004) 1645–1649.
[26] A. Mojsilovic, J. Gomes, B. Rogowitz, Isee: Perceptual features for image library navigation in: Proc. SPIE Human vision and electronic imaging, vol. 4662, San Jose, California, 2002, pp. 266–277.
[27] J. Vogel, Semantic Scene Modeling and Retrieval, no. 33 in Selected Readings in Vision and Graphics, Houghton Hartung-Gorre Verlag Konstanz, 2004.
[28] J. Vogel, B. Schiele, Natural scene retrieval based on a semantic modeling step, in: International Conference on Image and Video Retrieval, LNCS, vol. 3115, Dublin, Ireland, 2004, pp. 207–215.
[29] M. Varma, A. Zisserman, Texture classification: Are filter banks necessary?, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, Madison, Wisconsin, 2003, pp. 691–698.
[30] T. Leung, J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional textons, International Journal of Computer Vision 43 (1) (2001) 29–44.
[31] J. Portilla, E. Simoncelli, A parametric texture model based on joint statistics of complex wavelet coefficients, International Journal of Computer Vision 40 (1) (2000) 49–70.
[32] S. Lazebnik, C. Schmid, J. Ponce, A sparse texture representation using affine-invariant regions, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, Madison, Wisconsin, 2003, pp. 319–324.
[33] A. Bosch, A. Zisserman, X. Muñoz, Scene classification via plsa, in: European Conference on Computer Vision, vol. 4, Graz, Austria, 2006, pp. 517–530.
[34] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2005, pp. 524–531.
[35] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, L. Van Gool, Modeling scenes with local descriptors and latent aspects, in: International Conference on Computer Vision, Beijing, China, 2005, pp. 883–890.
[36] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, 2006, to appear.
[37] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Machine Learning 41 (2) (2001) 177–196.
[38] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
[39] Y. Teh, M. Jordan, M. Beal, D. Blei, Hierarchical dirichlet process, Neural Information Processing Systems 17 (2005) 1385–1392.
[40] B.C. Russell, A.A. Efros, J. Sivic, W.T. Freeman, A. Zisserman, Using multiple segmentations to discover objects and their extent in image collections, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, 2006, to appear.
[41] F. Perronin, C. Dance, G. Csurka, M. Bressan, Adapted vocabularies for generic visual categorization, in: European Conference on Computer Vision, vol. 4, Graz, Austria, 2006, pp. 464–475.
[42] A. Bosch, X. Muñoz, J. Marti, Using appearance and context for outdoor scene object classification, in: IEEE International Conference on Image Processing, vol. II, Genova, Italy, 2005, pp. 1218–1221.
[43] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from google's image search, in: International Conference on Computer Vision, vol. II, Beijing, China, 2005, pp. 1816–1823.
[44] A. Torralba, A. Oliva, Semantic organization of scenes using discriminant structural templates, in: International Conference on Computer Vision, Korfu, Greece, 1999, pp. 1253–1258.
[45] A. Oliva, A. Torralba, Scene-centered description from spatial envelope properties, in: International Workshop on Biologically Motivated Computer Vision, LNCS, vol. 2525, Tuebingen, Germany, 2002, pp. 263–272.
[46] A. Bosch, X. Muñoz, A. Oliver, R. Martí, Object and scene classification: what does a supervised approach provide us?, in: IAPR International Conference on Pattern Recognition, Hong Kong, 2006, to appear.
[47] M.D. Squire, W. Müller, H. Muller, T. Pun, Content-based query of image databases: inspirations from text retrieval, Pattern Recognition Letters 21 (2000) 1193–1198.
[48] J. Sivic, B. Russell, A. Efros, A. Zisserman, W.T. Freeman, Discovering objects and their locations in images, in: International Conference on Computer Vision, Beijing, China, 2005, pp. 370–377.

Please cite this article as: Anna Bosch, et al., A review: Which is the best way to organize/classify images by content?, Image and Vision Computing (2006), doi:10.1016/j.imavis.2006.07.015