Object and Scene Classification: what does a Supervised Approach Provide us?

Anna Bosch, Xavier Muñoz, Arnau Oliver, and Robert Martí
University of Girona
Campus Montilivi, Ed. P-IV, 17071, Girona
{aboschr, xmunoz, aoliver, marly}@eia.udg.es

Abstract

Given a set of images of scenes containing different object categories (e.g. grass, roads) our objective is to discover these objects in each image, and to use this object occurrences to perform a scene classification (e.g. beach scene, mountain scene). We achieve this by using a supervised learning algorithm able to learn with few images to facilitate the user task. We use a probabilistic model to recognise the objects and further we classify the scene based on their object occurrences. Experimental results are shown and evaluated to prove the validity of our proposal. Object recognition performance is compared to the approaches of He et al. [3] and Martí et al. [6] using their own datasets. Furthermore an unsupervised method is implemented in order to evaluate the advantages and disadvantages of our supervised classification approach versus an unsupervised one.

1. Introduction

Classifying scenes (such as mountains, forests, offices) is not an easy task owing to their variability, ambiguity, and the wide range of illumination and scale conditions that may apply. Several studies suggest 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 [10]. Hence our aim is to first recognise objects in images, and then use them to classify the scenes.

Several works on that topic have been proposed in the literature. For instance in [12], they divided the image in a 10×10 grid and classified each patch as a certain object (e.g. sky, grass, etc.) referred to as *local semantic concepts*. Then, each image is described by the frequency of occurrence of a semantic concept and they construct a prototype of each scene category. A drawback of this method is the learning cost and the fact that more than one object can be found in a patch. In [5] they used a semi-supervised approach which reduces the learning cost. The system learns an intermediate representation, called *themes* and the image clasification is performed according to the theme distribution. Also in [2, 9] image classification is based on an intermediate representation, using unsupervised latent space models. In this case, a probabilistic Latent Semantic Analysis (pLSA) is used to generate a compact scene representation and classify the images based on this representation. A common issue in cited approaches is that they can not label objects in images as humans do, they find and represent different parts of the images sometimes corresponding to parts of the same object.

Our object and scene classification proposal tries to improve the above methods in two ways. First we recognise and represent the objects and the whole image similarly to human perception. This means that we recognise the objects in images and not patches or parts of them. The main apportation here is the classification of scenes based on their *segmented and recognised* objects. The second novel aspect is the use of a small number of images in the supervised learning process. Hence we use a supervised learning with few images and involving the user as less as possible.

This paper is structured as follows. Section 2 describes how the system achieves the object recognition and scene classification. In Sections 3 and 4 we explain the datasets and methodology used to test our approach. Following, Section 5 shows the results obtained and a comparison with other object and scene classification methods. A final discussion and conclusions are given in Section 6.

2. System Overview

2.1. Object Recognition

In this Section we briefly describe our supervised object recognition method (see [1] for a further explanation). A basic schema of the system is shown in Figure 1 and works as follows:

Learning: The learning is carried out by using a small number of images to train the system, obtaining a simple and 'general' initial model for each object, which contains its appearance and contextual position. The learning carries out a feature selection process to select for each single object the specific subset of features which best differentiates the current object from the rest.

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

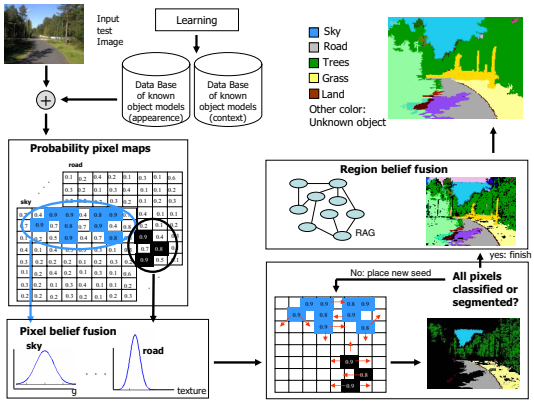

Figure 1. Proposed hybrid method for the classification and segmentation of the image.

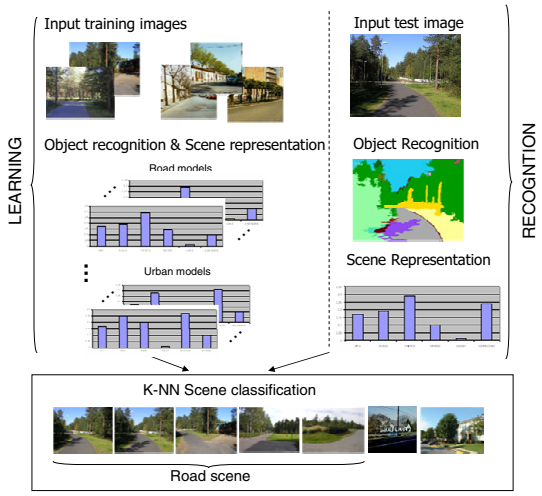

Figure 2. Proposed scene representation and classification methodology.

Recognition: The recognition process uses the knowledge of the learned objects to obtain the probability of every pixel of belonging to each object, obtaining probabilistic pixel maps (one map for each object). The appearance probability of a pixel j characterised by the features $\overrightarrow{x_j}$ of belonging to an object $\emptyset_i$ is given, under a Gaussian assumption, by the following probability density function:

$$P_A(j|\emptyset_i) = \frac{1}{\sqrt{(2\pi)^k|\Sigma_i|}} \exp\{-\frac{1}{2}(\overrightarrow{x_j}-\overrightarrow{\mu_i})^T\Sigma_i^{-1}(\overrightarrow{x_j}-\overrightarrow{\mu_i})\} \quad (1)$$

where $\overrightarrow{\mu_i}$ is the mean feature vector of the object $\emptyset_i$, Σ_i its covariance matrix, and k the number of characteristics. In addition, we express the probability that a pixel j at position y_j belongs to an object $\emptyset_i$ considering its absolute position by the following equation:

$$P_L(j|\emptyset_i) = max(L_{T_i} * P_T(y_j), L_{M_i} * P_M(y_j), L_{B_i} * P_B(y_j)) \quad (2)$$

where L_{T_i}, L_{M_i}, and L_{B_i} are the learned probabilities for object $\emptyset_i$ to be in the top middle and bottom of the image respectively and $P_T(y_j)$, $P_M(y_j)$ and $P_B(y_j)$, are the believes that a pixel j with y position is to a certain location (top, middle, bottom) in the image. Therefore, the merging of both probabilities, P_R, provides a probabilistic pixel map for each object:

$$P_R(j|\emptyset_{Li}) = P_A(j|\emptyset_{Li}) * P_L(j|\emptyset_{Li}) \quad (3)$$

The main contribution in our approach lies in the next stage: the most probable pixels of each map are detected, and constitute the core of the objects. Those pixels are used as samples to extract a new and more accurate model which uses as object characteristics the information given by the core pixels of the current test image. The posterior growing of *specific active regions* from these cores allows to classify and segment the image. Here we use specific and different features (the ones obtained by a feature selection process) to grow each object. Until here the algorithm follows a top-down step, since the knowledge is used at the beginning of the process. However, the next stage is a bottom-up control applied by performing a general purpose segmentation of unclassified areas, which allows us to extract the unknown objects without any previous information of them. Finally, a last stage of region belief fusion exploits the contextual information provided by neighbouring objects to refine the initial classification of unknown regions.

This approach allows to recognise and segment the objects similarly as humans do. We are able to label regions instead of patches, which provide a more accurate classification of the local semantic concepts. Moreover, we are using a supervised learning approach which learns with a few images, so the learning stage is not expensive in terms of human interaction. The supervised learning provides another important feature: the system knows which label has to assign to each object. This conforms an important difference compared to the systems proposed in [5, 9] where the intermediate representations do not correspond with objects that humans perceive.

2.2. Scene Classification

After the *semantic classification* (object recognition) we want to classify scenes represented by semantic modeling. In the literature we can find several proposals that generate a model for each scene [5, 12] while others have several models per scene category [8]. We propose to have several prototypes for each kind of scene, specifically, we have a model for each training image. This model is acquired computing the probability of each object to be in an image. In other words, for each image we will have a vector (called M) with the probabilities of the occurrence of objects. Thus we will represent each image as:

$$M_I = [P_{\emptyset_1}, P_{\emptyset_2}, ..., P_{\emptyset_N}] \quad (4)$$

where $P_{\emptyset_i}$ is the probability of $\emptyset_i$ to appear in the image

Figure 3. Images from OU and MA datasets: (a) road, (b) suburb, and (c) city scenes.

and is computed using the following equation:

$$P_{\emptyset_i} = \frac{1}{N_p} * Obj(I, \emptyset_i) \quad (5)$$

where N_p is the total number of pixels in the image I and $Obj(I, \emptyset_i)$ is the total number of pixels in image I recognised as object $\emptyset_i$.

The result is that each image is represented by an N-vector where N is the number of objects. To classify a test image as a certain scene category, we first recognise the objects and construct its object representation. The test image is then classified using a K-Nearest Neighbours classifier (KNN) on the N-vectors of the training images. An Euclidean distance function is used. In more detail, the KNN selects the K nearest neighbours of the new image within the training database. Then it assigns to the new picture the label of the closest category which is most represented within the K nearest neighbours. Figure 2 shows graphically the learning and scene classification process.

3. Datasets

We evaluated our classification algorithm using two different datasets: (i) Outex dataset [7], and (ii) Martí et al. dataset [6]. We will refer to these datasets as OU and MA respectively. These images consist of natural outdoor scenes and mainly contain typical objects in rural and suburban areas. Figure 3 shows example images from each dataset, and the contents are summarised here:

OU: includes 41 images of natural outdoor scenes. The average size of each image is 256×192 pixels.

MA: includes 87 natural scenes taken by themselves. The size of the images is 250×250 or 204×137. Every scene category is characterised by a high degree of diversity of meteorological conditions and different seasons of the year.

We merged these two datasets (obtaining 128 images) and organised their images into three different scene categories: 43 *road*, 43 *suburb* and 42 *city*. We segmented and labeled them manually into 7 objects: *sky*, *grass*, *road*, *vegetation*, *dark house*, *white house* and *ground*, while the remaining areas, mainly belonging to man-made objects, are considered as *unknown* objects.

4. Evaluation Methodology

In order to evaluate the goodness of the implemented system a comparison between the results of the classifications system and hand-labeled objects/scenes is performed. Specifically, to know the performance of the system a confusion matrix is computed. The overall performance rates are measured by the average value of the diagonal entries of the confusion matrix. We randomly selected 35 training images and the remaining ones (93 images) are used for testing. This number of training images was stated in our experiments as a good compromise between the required effortless of the user and the quality of results.

The object recognition system is compared to the works of Martí et al. [6], He et al. [3], Bosch et al. [2], and to a baseline method in order to gauge the difficulty of this task:
Pixel-based classifier. Every image pixel is classified as the object with the highest appearance probability P_A (see Equation 1) always this is higher than a fixed threshold ($P_A > 0.7$). Otherwise, the pixel is labeled as unknown. No context information is used.

The scene classification scheme is also compared to the work in [2] and to a baseline method:
Global colour model. The algorithm computes global HSV histograms for each training image. The colour values are represented by a histogram with 36 bins for *H*, 32 bins for *S*, and 36 bins for *V*, giving a 84-dimensional vector for each image. A test image is classified using KNN (with $K = 6$).

5. Experimental Results

Implementation: Once the user has selected the object the system extracts the features of each pixel contained in the selected area. For colour information, we use the RGB, HLS and CIE Lab* colour spaces, the later being perceptually uniform. The texture information is obtained by a set of co-occurrence matrix-based features by using a distance of one pixel and angles quantised to 45º intervals. Hence, four matrices of horizontal, first diagonal, vertical, and second diagonal (0º, 45º, 90º, 135º) are used. The statistics applied were: Contrast, Homogeneity, Correlation and Entropy. Thus each pixel is represented by 25 image statistics. We used $K = 6$ for the KNN in scene classification.

The learning stage takes approximately 45 minutes, considering the feature selection and without taking the time used to select the training images into account. On the other hand, the classification task takes about 2 hours for testing images (Visual C++ and php implementation on a 1.7GHz PC).

5.1. Object recognition

We obtain a rate of approximately 87% when classifying 7 objects in images using the described method, while

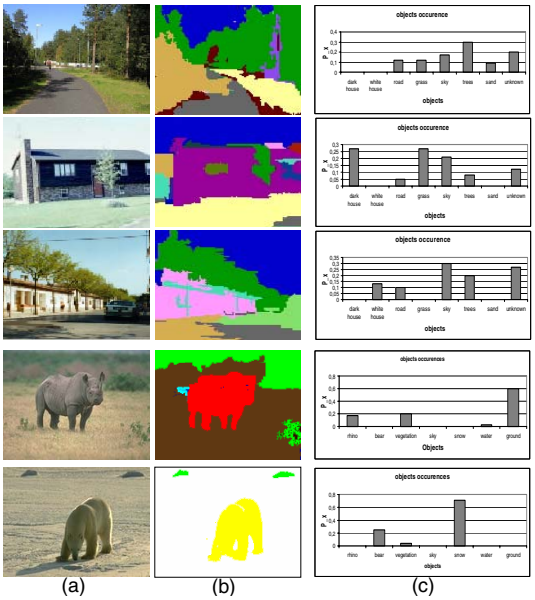

Figure 4. Results when classifying objects by the supervised method: (a) original image, (b) object recognition, (c) object occurrence.

a 67% is obtained using the pixel-based classifier method. This means that the use of data from the test image and context information during the training stage is important for a better performance. Figure 4 shows some qualitative results. We can see in this Figure that unknown objects (tree shadows at first row, windows at second row and shadows on the road at third row) are not recognised and the system paints them with a different colour.

Authors in [6] obtained a 87% (including training images in the classification process) using the same MA dataset when recognising four objects: *sky*, *leaves*, *road*, *ground*. In order to carry out with a comparison, we used the MA dataset and tried to recognise only these four objects. We used 20 training images and the remaining 67 for testing and the score obtained was 90%.

In order to compare the goodness of our method when working with other kind of images, we tested it with the same dataset used in [3] (HE dataset). It is a 100 image subset of the Corel image database, consisting of African and Arctic wildlife natural scenes. The hand labeled images were provided by the authors of the paper. Each image is 180×120 pixels. They manually labeled them into 7 classes: *rhino/hippo*, *polar bear*, *vegetation*, *sky*, *water*, *snow* and *ground*. They did not take unknown objects in the images into account, so all the regions in these images are known. They obtained an score of 80% of correct classified objects using a multi Conditional Random Field. We outperform this result to 86.76% with the same images. Moreover, they used 60 training images while in our experiments the training set was composed by 30 images. Qualitative results and object occurrence in two images from this dataset are shown at last two rows in Figure 4.

Table 1. Overall rates for object classification

Database	#obj.	% Proposal	% Other	Other #Ref.
OU & MA	7	87%	69%	Baseline Section 4
OU & MA	7	87%	53%	Bosch et al. [2]
MA	4	90%	87%	Martí et al. [6]
HE	7	86%	80%	He et al. [3]

Furthermore, in [2] we presented an unsupervised method to discover *topics* ('objects') in images. This method was based on pLSA, a generative model from the statistical text literature [4]. This method was applied here over the same dataset than the supervised one (OU & MA), and 15 topics were discovered. We obtained a 53% of correctly classified objects. Table 1 summarises the overall rates obtained with our object classification proposal and the comparative evaluation.

5.2. Scene classification

Table 2 shows the confusion matrix when classifying images into one of the three predefined scenes. The average score is 79.8%. Most confused scene is *city* with *suburb*. This could be explained by the fact that the same objects are present in most of the images belonging to these two scenes. Hence these recognised objects are not enough to correctly classify the scene. The baseline algorithm obtains a 62% when classifying these images.

Observing in more detail the three kinds of scenes we can extract the following conclusions: (i) all the three scenes contain the objects *sky*, *trees* and *road*, (ii) *grass* object mainly appears in *road* and *suburb* scenes. (iii) *houses* are distinctive of *suburb* and *city*. Hence it is not an easy task to separate these scenes from their objects, even for humans. This is extensively discussed in Section 6.

Figure 4c shows the object occurence for one image of each scene. Note that they are very similar, showing the difficulty to classify these images using these objects. This could be due to the fact that we do not have only one model to represent each scene but several ones. Then if a test image is similar to a training one, their object occurrence will be almost the same and will be well classified. On the other hand, the object occurrence for the scenes of the last two rows in Figure 4 (*African* and *Artic* scene from HE dataset) would be easily separable, since their object occurrences are very different.

In [2] the unsupervised approach was extended to a semi-supervised scene classification. An image containing instances of several objects is modeled as a mixture of topics (each image is also represented as a vector). The test image is then classified using a KNN ($K = 3$) on the vectors of the training images. A rate of 96% is obtained when classifying the three scenes using the leave-one-out methodology.

Table 2. Confusion matrix when scene classification

		CLASSIFICATION		
		road	suburb	city
TRUTH	road	81.0%	9.0%	10.0%
	suburb	6.5%	79.5%	14.0%
	city	0.0%	21.0%	79.0%

6. Summary & Discussion

We demonstrated that our supervised object recognition system is able to learn the same objects similarly to humans giving very good results when classifying 7 objects. It is able to give a label for each object and also to identify and segment the unknown ones. Moreover, we compared it to two previous supervised natural object classifiers [3, 6] and we achieved better results using their own datasets and fewer training images. Obviously the unsupervised method [2] is better from the point of view of the user interaction. However, the proposed system learns with very few images per object category compared to previous supervised approaches. Moreover far better rates of object classification are achieved with the supervised compared to the unsupervised method. This is because we teach and show the system how the objects are from our point of view, then, it is able to learn object perception similarly to humans. In contrast, the unsupervised method is free an learns the topic representation by its own way. This involves that for the system, a *blue sky* would be the same object than a *blue car* and the system labels two different objects as the same.

On the other hand, when classifying scenes the unsupervised approach is better. This is because the unsupervised has the freedom for choosing topics, the system organises them in their own way in order to have different object representation for the different scenes. For example it distinguishes the sky as three different objects: *blue sky*, *grey sky* and *sky with clouds*. Observing the MA and OU datasets, we can see that most of the road scenes have a blue sky, most of the suburb scenes have a grey sky, while sky in city scenes mostly contains clouds. It gives us additional information to classify scenes because image representation is more discriminative (note that the number of topics is higher (15) than when using the supervised method (7)). Comparing three first rows in Figure 4 with images in Figure 5 we can see that objects are best labeled when using the supervised approach but objects occurrences (or topic distributions) are more discriminative to classify scenes when using the unsupervised one.

Although we only classify 3 scene classes, the task is not easy, since they are very similar and ambiguous taking their semantic content into account. Moreover, this is a more ambitious task than classical *indoor/outdoor* scene classification. These cases are very easily separable due to the different objects and structure. In addition, the set of images and categories used by authors are often constrained. As an example, the categories used by Vailaya et al. 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". After demonstrating the good performance of our approach, further exploration will focus on the combination of supervised and unsupervised techniques to take advantage of both approaches.

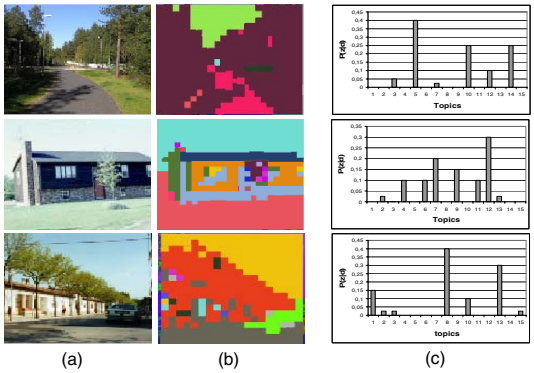

Figure 5. Results when classifying objects by the unsupervised method: (a) original image, (b) object recognition, (c) topic distribution.

Acknowledgements

Thanks to X. He and J. Freixenet for providing their datasets. This research was sponsored by the grant BR03/01 from the University of Girona.

References

[1] A. Bosch, X. Muñoz, and J. Marti. Using appearance and context for outdoor scene object classification. In *ICIP*, 2005.

[2] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via plsa. In *ECCV*, 2006.

[3] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[4] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *ML*, 41(2):177–196, 2001.

[5] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[6] J. Martí, J. Freixenet, J. Batlle, and A. Casals. A new approach to outdoor scene description based on learning and top-down segmentation. *IVC*, 19:1041–1055, 2001.

[7] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, and S. Huovinen. Outex - new framework for empirical evaluation of texture analysis algorithms. In *ICPR*, 2002.

[8] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175.

[9] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV*, 2005.

[10] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

[11] A. Vailaya, A. Vigueiredo, A. Jain, and H. Zhang. Content-based hierarchical classification of vacation images. In *ICMCS*, 1999.

[12] J. Vogel. Semantic Scene Modeling and Retrieval, volume 33 of *Selected Readings in Vision and Graphics*. Houghton Hartung-Gorre Verlag Konstanz, 2004.