Available online at www.sciencedirect.com
ScienceDirect
Image and Vision Computing 25 (2007) 727–740
www.elsevier.com/locate/imavis

Segmentation and description of natural outdoor scenes

A. Bosch *, X. Muñoz, J. Freixenet

University of Girona, Institute of Informatics and Applications, Campus Montilivi, Edifici P4, Av. Lluis Santaló, s/n, 17071-Girona, Spain

Received 24 November 2005; received in revised form 7 March 2006; accepted 16 May 2006

Abstract

A scene description and segmentation system capable of recognising natural objects (e.g. sky, trees, grass) under different outdoor conditions is presented. We propose an hybrid and probabilistic classifier of image regions as a first step in solving the problem of scene context generation. We focus our work in the problem of image regions labeling to classify every pixel of a given image into one of several predefined classes. The result is both a segmentation of the image and a recognition of each segment as a given object class or as an unknown segmented object. Classification performance has been evaluated with the Outex dataset and compared to the approach of Martí et al. (IVC 2001) and He et al. (CVPR 2004) using their own datasets, showing the superiority of our method.

Keywords: Image understanding; Object classification; Object segmentation

1. Introduction

We tackle in this paper the problem of natural object labeling to classify every pixel of a given image into one of several predefined classes. Hence, we might consider images of outdoor scenes and we would like to classify each pixel as *sea*, *snow*, *road*, etc. To achieve this goal, and in the absence of any prior information, the scene classification task requires the knowledge of objects contained in the image. There are a lot of researchers that assume as knowledge only the appearance of objects (colour, texture and shape). As recent examples, Vailaya et al. [1] used spatial colour moment and edge direction histograms in order to classify scene categories, such as indoor/outdoor, city/landscape. Barnard et al. [2,3] considered colour, texture and shape information to solve the object recognition problem, and similar descriptors are used in [4] to generate maps segmented into objects of interest: buildings, vegetation and so on. Li et al. [5] have used gray level patches and SIFT features to classify themes in natural scenes without supervision. SIFT features are also used in [6,7]. Nevertheless, it is increasingly being recognised in the vision community that context information is necessary for a reliable extraction of the image regions and objects [8–11].

Another important issue in image understanding is the overall control of the system (in which step will we use the knowledge information acquired during the learning?). Batlle et al. [12] describe three types of hierarchical control: top-down [13–15], bottom-up [16–18] and hybrid [4,19,20]. The first of them, can be described as *hypothesise-and-test*, once a hypothesis is generated it uses the knowledge acquired at the learning stage to verify the hypothesised object. This approach is limited by its inability to handle unexpected regions (corresponding to unknown objects), but can handle variations, exceptions and special cases that are known a priori. On the other hand, bottom-up systems follow an opposite approach, they do not use the knowledge at the low level image processing stages, which are mainly based on a general purpose image segmentation. Hence, these kind of systems are much better at handling unexpected regions than those using a top-down strategy, therefore they even would be able to provide a description of unknown objects found in the image. Finally, hybrid approaches seek to get the best of both approaches.

* Corresponding author. Tel.: +34 972418891; fax: +34 972418098.
E-mail addresses: aboschr@eia.udg.es (A. Bosch), xmunoz@eia.udg.es (X. Muñoz), jordif@eia.udg.es (J. Freixenet).

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2006.05.015

Our previous approach [13] (proposed in 2001) was a top-down control system which labeled every pixel in the image as a certain object (e.g. leaves, road). The technique achieved successful results and had two main characteristics to remark: (i) the facility to teach the system, providing a very easy and intuitive interface; and (ii) the way which the system deals with different outdoor conditions. However it had also some drawbacks: first, the system takes advantage of a top-down approach to recognise learned objects in the image. Nevertheless, it is not able neither to handle nor give information about segmented unknown objects. Second, it used a set of discriminative classifiers (a decision tree for each object) to label each pixel. Thus, a pixel can be labeled as two different objects without knowing which one is the best. Third, the system is not capable to improve results when initial over-segmentation (it cannot rectify initial wrong labels).

The proposed method solves the drawbacks mentioned above. We propose a probabilistic classifier (taking appearance and contextual information into account) to recognise regions belonging to scenes primarily containing natural objects. Furthermore, the technique handles with known and unknown objects in the image by following an hybrid control. The approach is inspired in particular by three previous papers: (i) using information about the learned models and also information provided from the test image to perform the classification [18]; (ii) the use of active regions to perform the classification [21] and take the neighbours of a pixel into account; and (iii) the use of a supervised learning with a very intuitive interface to acquire the knowledge about objects [13] taking the different seasons and meteorology conditions into account. We have made extensions over all three of these works as we will show in the rest of the paper.

This paper is organised as follows. Section 2 describes our proposal, taking the phase of learning and recognition into account. In Sections 3 and 4, we explain the used datasets to evaluate the system and we give the implementation details: features used and value of parameters, as well as an explanation of the methodology used to evaluate the system performance. In Section 5, some experimental results are shown and discussed in Section 6. We evaluate the performance of our system and its ability to handle with known and unknown objects. Moreover, the results are compared with the results of [13] and the results of a recent work [22] with their own datasets. We finish the paper with the conclusions and some ideas of further work.

2. System overview

Three questions have to be addressed in order to pursue our idea: How to use the learning information? How to obtain the classification and segmentation of the known and unknown objects of the test image? How to use contextual information? In this section, we address these questions in a Bayesian setting and by an *specific* active region-based segmentation.

We propose to solve these questions by using few images to train the system obtaining a simple and 'general' initial model for each object, which contains its appearance and absolute context. The learning carries out a feature selection process to chose for each single object the specific subset of features which best differentiate the current object from the remaining ones. The recognition process starts by using the knowledge of the learned objects to obtain the probability of each pixel to belong to each object. This provides us a set of probabilistic pixel maps (one map for each object). The most probable pixels of each map are detected, and are going to constitute the core of objects (CO). The COs are used to extract a new and more accurate object model. The posterior growing of *specific* active regions from these CO allows us to classify and segment the image. Until here the algorithm follows a top-down control, since the knowledge is used at the beginning of the process. Following, a bottom-up control is applied to perform a general purpose segmentation of not-classified areas. This extracts 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. Fig. 1 shows the basic architecture of our proposal, and each stage is described at the following sections.

2.1. Learning

In recent but conventional schemes [18,22–25], a model for each class is learnt using a lot of training data belonging to that class and a probability based on the learnt model is assigned to the newly observed data. However, it is possible that the input test image has been generated from a subset of the full generative model support, and using the full model to assign generative probabilities can produce serious artifacts in the probability assignments. In [18] they proposed a method to constraint the overall model using the distribution of the newly observed data.

Learning a new object for humans is fast and easy, sometimes requiring very few training examples in contrast to above approaches. The proposal of [26–28] is able to learn object categories (e.g. cars, airplanes) from just few images. The advantage of using few images is that the stage of learning is not so hard, and the assisted steps are reduced which alleviates the user of an habitually tedious and expensive task. On the other hand, the main drawback of using few images in natural images is that the results may not be satisfactory, because not the whole region of the model can be found. Hence our goal at this stage is to perform the learning by using few images, and to consider how to model the information to carry out the later recognition.

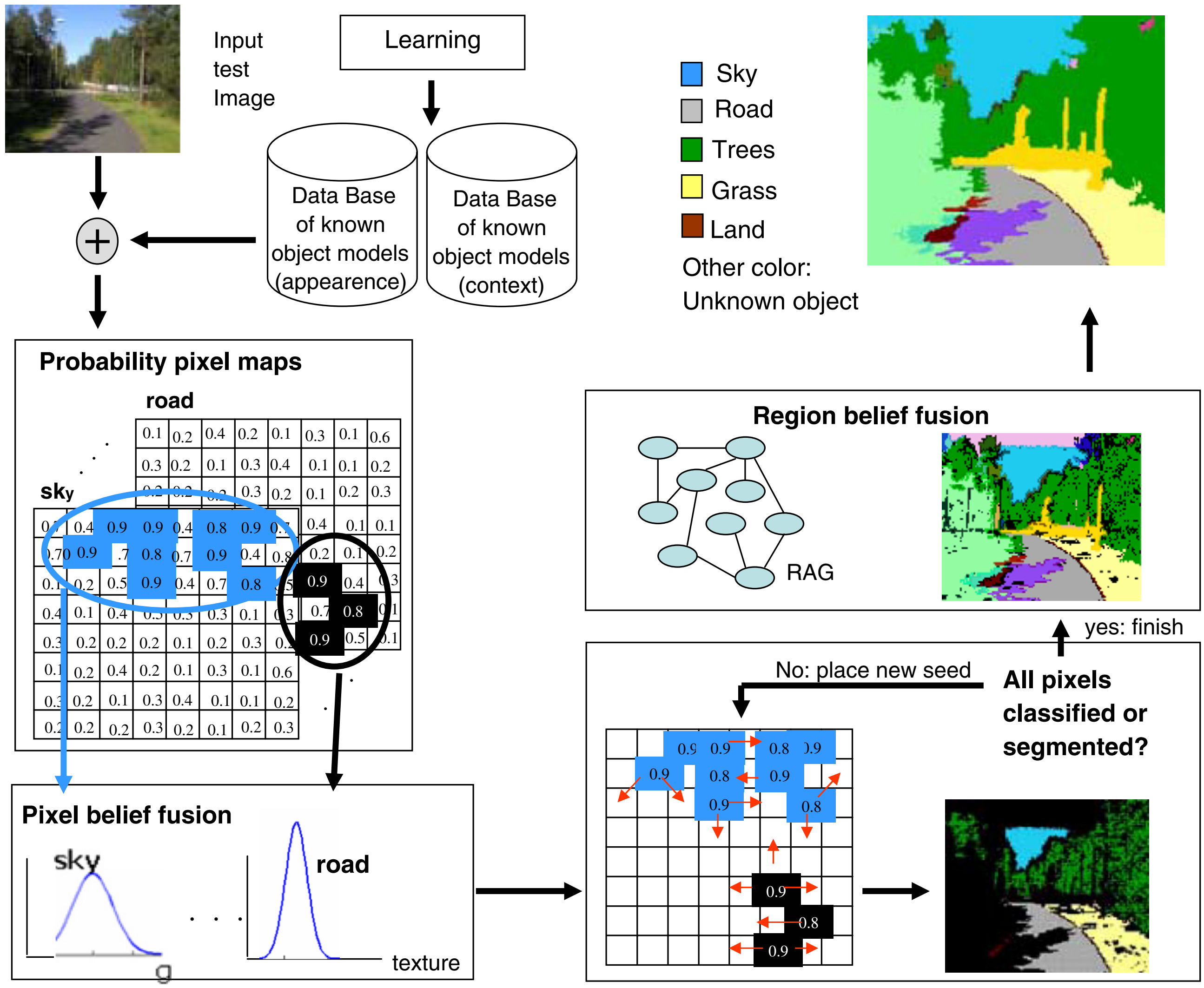

Fig. 1. Proposed hybrid method for the classification and segmentation of the image.

The learning stage consists of three phases which are shown in Fig. 2 and explained in the following sections:

2.1.1. Select the training images

The user must select a set of training images that contain all the objects that he/she wants to teach to the system. This allow us to learn new categories from few training examples. In order to handle with objects in different weather conditions, we model each object as a set of *object classes*. An *object class* is a prototype of a real object described in terms of colour and textural features under specific outdoor conditions like in [13]. Thus, we will have the object *sky* with, for example, three object classes: cloudy sky, sunny sky and storm sky.

2.1.2. Feature extraction

The objects to be learnt must be shown to the system. We developed a very easy and intuitive web interface to carry out this stage, where the user selects meaningful examples of objects by drawing a square on the object of interest (see Fig. 2). From this selected area, a set of pixels is extracted and considered as samples of the object. Next, a large number of colour and texture features are measured (see Section 4.1 for more details about the features used).

Besides, the system learns the absolute contextual information of the object to know where the object is generally placed in the image (see Fig. 2). This information is computed by using a vote score: we split the image into three horizontal areas, and we consider that each object could be located at *top* if $y_j \in [0, Y_T)$ (where y_j means the y position of pixel j), *middle* if $y_j \in [Y_T, Y_B)$ or *bottom* of the image if $y_j \in [Y_B, Y_{SIZE}]$. Note that 0 is at top of the image and Y_{SIZE} at the bottom. Then every time that the system learns a new object sample, it computes where it is located, and increments the corresponding position score of the object. Following this, each object location is represented by a normalised three-dimensionality vector (the positions represent *top*, *middle* and *bottom*). For instance, the location of *sky* object will be represented by (1,0,0), which means that it is always at top position, and never at middle or bottom.

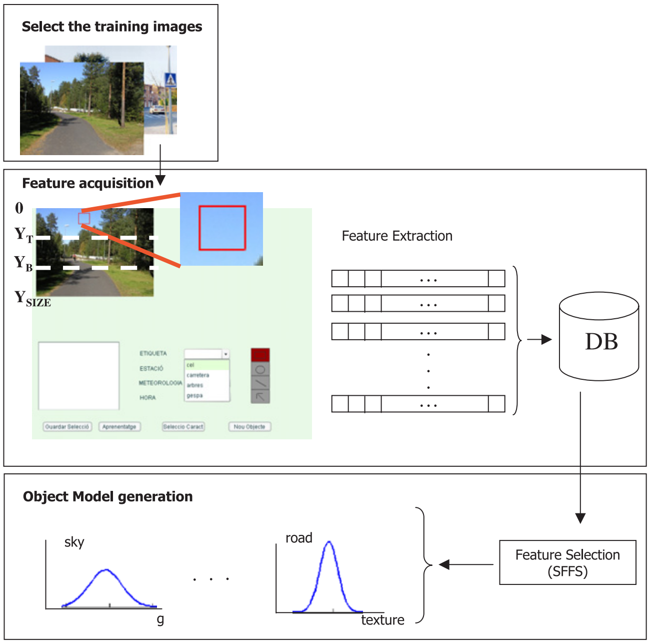

Fig. 2. Three phases for the learning stage of our system: (i) select the training images; (ii) feature extraction; and (iii) object model generation.

2.1.3. Object model generation

From these samples, a set of initial objects (Ø_I) with their characteristics vectors $\text{Ø}_I = [\text{Ø}_{I1}(\overrightarrow{f_1}, \ldots, \overrightarrow{f_n}), \ldots, \text{Ø}_{Ik}(\overrightarrow{f_1}, \ldots, \overrightarrow{f_n})]$ is computed. Nevertheless, it is well known that using the whole set of features does not always mean to improve the quality of classification. Moreover, due to the complexity of outdoor scenes, it is necessary to emphasise that not all object classes are defined in terms of the same attributes. Consequently, every single object can be described by specific features, in order to facilitate the characterisation and later recognition process, and improve the accuracy classification. The system performs a feature selection process for each single object, with the goal to find the subset of features which best differentiates the current object to the remaining ones. A classical Sequential Forward Floating Search (SFFS) algorithm [29] is used with this aim. As a result, a new set of objects with fewer characteristics called Ø_F is obtained: $\text{Ø}_F = [\text{Ø}_{F1}(\overrightarrow{f_p}, \ldots, \overrightarrow{f_q}), \ldots, \text{Ø}_{Fk}(\overrightarrow{f_l}, \ldots, \overrightarrow{f_m})]$ where $0 < p, q, l, m \leqslant n$. Next, considering the selected features, we assume that each object is modeled by a Gaussian distribution characterised by $\overrightarrow{\mu_i}$ (the mean vector of the object i) and Σ_i (its covariance matrix). Hence, the final learned object model set can be defined as: $\text{Ø}_L = [\text{Ø}_{L1}(\overrightarrow{\mu_1}, \Sigma_1), \ldots, \text{Ø}_{Lk}(\overrightarrow{\mu_k}, \Sigma_k)]$.

2.2. Segmentation and classification

Recognition of objects is performed by using the models acquired on the previous learning. This initial knowledge is used to obtain a probabilistic pixel map for each object, and also a first classification. However, we consider this pixel-level classification only as a first step in the recognition process with the aim to initiate the object recognition by *specific* active region segmentation. The inclusion of a higher region-level information allows the system to take into account the spatial consistency of objects in the image, which highly improves the classification accuracy [4].

2.2.1. Probabilistic pixel map

The system starts by an initial classification of image pixels in order to obtain a set of probability maps. Each map is associated to a known object and contains the probability for every pixel of the test image to be classified as the current object. We use the models acquired from the learning to calculate the probability that a pixel belongs to an object.

The appearance probability of a pixel j characterised by the features $\overrightarrow{x_j}$ of belonging to a object Ø_{Li} is given, under a Gaussian assumption [30], by the probability density function:

$$P_A(j|\text{Ø}_{Li}) = \frac{1}{\sqrt{(2\pi)^k|\Sigma_i|}} \exp\{-\tfrac{1}{2}(\overrightarrow{x_j} - \overrightarrow{\mu_i})^{\mathrm{T}}\Sigma_i^{-1}(\overrightarrow{x_j} - \overrightarrow{\mu_i})\} \tag{1}$$

where $\overrightarrow{\mu_i}$ is the mean vector of the object Ø_{Li}, Σ_i its covariance matrix, and k the number of characteristics. Note that k is also the dimensionality of $\overrightarrow{x_j}$ and $\overrightarrow{\mu_i}$ vectors. This dimension is the number of features used to represent each object and its value depends on the feature selection process (see Section 2.1.3 for the feature selection process and Section 4.1 for the features which best represent each object).

At this stage, we compute a contextual probability by using a fuzzy rule based approach (see Section 4.1 for the fuzzy rules implementation). For each object we learned its habitual location in the image, which is described by the percentages of being at the *top*, *middle* and *bottom* of an image, (L_{T_i}, L_{M_i}, and L_{B_i}, respectively). Now, at the recognition stage, the y position of all pixels is obtained and the probability of each of them to belong to a certain position is computed. Fig. 3 shows the fuzzy rules used to provide the position of pixels in a fuzzy way. The probabilities $P_T(y_j)$, $P_M(y_j)$ and $P_B(y_j)$, are the belief that a pixel with y_j position is to a certain location (top, middle, bottom) in the image. Therefore, Eq. (2) gives us the probability that a pixel j at position y_j belongs to an object Ø_{Li} considering its absolute position:

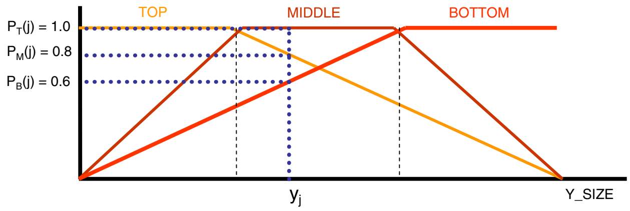

Fig. 3. Fuzzy rules for the initial context information, which provide the position of a pixel in the image. The origin 0 of Y_Size is considered at the top of the image.

$$P_L(j|\text{Ø}_{Li}) = \max(L_{T_i} * P_T(y_j),\ L_{M_i} * P_M(y_j),\ L_{B_i} * P_B(y_j)) \tag{2}$$

This kind of contextual information is useful at this initial stage in order to differentiate objects with similar appearance but different locations, such as white clouds and the snow, and to avoid its confusion. Therefore, the merging of both probabilities, P_R (Eq. (3)), provides a probabilistic pixel map for each object.

$$P_R(j|\text{Ø}_{Li}) = P_A(j|\text{Ø}_{Li}) * P_L(j|\text{Ø}_{Li}) \tag{3}$$

2.2.2. Pixel belief fusion

Nevertheless, there are only a few pixels with a very high probability to belong to a certain object, so a reduced set of pixels can be classified at this time, with a high confidence of taking the right decision. This is due to the fact that few images have been used in the learning stage to construct the initial object models, and specially because objects in outdoor images have a really high variability, which implies the possibility of important differences between the learnt object and the one we are trying to recognise.

Inspired by the proposal of [18] we can improve the initial objects model by using the distribution of the newly observed data. The pixels with the highest probability to belong to an object ($P_R > 0.8$) constitute the CO, and are considered as representative data to design a lesser constrained new model. To construct this new model, for each object, the same features found in the previous object-specific feature selection process are taken into account, but $\overrightarrow{\mu_i}$ and Σ_i, which characterises the model, are re-computed (using the test data information) so the model represents the reality of the test image. This new set of objects is called $\text{Ø}_N : \text{Ø}_N = [\text{Ø}_{N1}(\overrightarrow{\mu_1}, \Sigma_1), \ldots, \text{Ø}_{Nk}(\overrightarrow{\mu_k}, \Sigma_k)]$.

2.2.3. Object belief refinement

The core pixels are used as starting seeds to initialise the growing of a concurrent set of *specific* active regions. Regions start to grow from the core pixels guided by their specific object model, as the colour and texture image data in order to segment the whole object based on minimising a global energy function. A similar technique was used in [21] to perform image segmentation. We improved that work in two ways: (i) they used the same features for each model to segment the images (active region segmentation) while we use specific features in order to segment each specific object (*specific* active region segmentation), providing a more accurate result; and (ii) we classify every segmented region as one of the learned object.

With the aim of integrating region and boundary information in an optimal segmentation/classification and to obtain an accurate result, the global energy is defined with two basic terms (see Eq. (5)). The region energy term measures the homogeneity in the interior of the regions by the probability that these pixels belong to each corresponding object using its specific features. The probability P_R is used to compute the region homogeneity. Meanwhile, the boundary term measures the probability that boundary pixels are really edge pixels. Nevertheless, it is well known that the extraction of accurate boundary information on textured images is a very tough task. We shall consider that a pixel j constitutes a boundary between two adjacent regions, A and B, when the properties at both sides of the pixel are different and fit with the models of both objects. Textural, colour and location features are computed at both sides (referred to as m and its opposite as n). Therefore, $P_R(m|\text{Ø}_A)$ is the probability that features obtained on the side m belong to object A, while $P_R(n|\text{Ø}_B)$ is the probability that the side n corresponds to object B. Hence, the probability that the considered pixel is boundary between A and B is equal to $P_R(m|\text{Ø}_A) \times P_R(n|\text{Ø}_B)$, which is maximum when j is exactly the edge between objects A and B as both sides obtain the better fit for both models. Four possible neighbourhood partitions (vertical, horizontal and two diagonals) are considered as in the proposal of [31] (see Fig. 4). Therefore, the corresponding probability of a pixel j to be boundary, $P_B(j)$, is the maximum probability obtained on the four possible partitions and $P_B(j|A,B)$ is defined as in Eq. (4).

$$P_B(j|A,B) = P_R(m|\text{Ø}_A) * P_R(n|\text{Ø}_B) \tag{4}$$

Some complementary definitions are required: let $\rho(R) = \{R_i : i \in [0,N]\}$ be a partition of the image into $N+1$ non-overlapping regions, where R_0 is the region corresponding to the background region. Let $\partial\rho(R) = \{\partial R_i : i \in [1,N]\}$ be the region boundaries of the partition $\rho(R)$. The energy function is defined as

$$E(\rho(R)) = (1-\alpha)\sum_{i=1}^{N} -\log P_B(j : j \in \partial R_i) + \alpha \sum_{i=0}^{N} -\log P_R(j : j \in R_i) \tag{5}$$

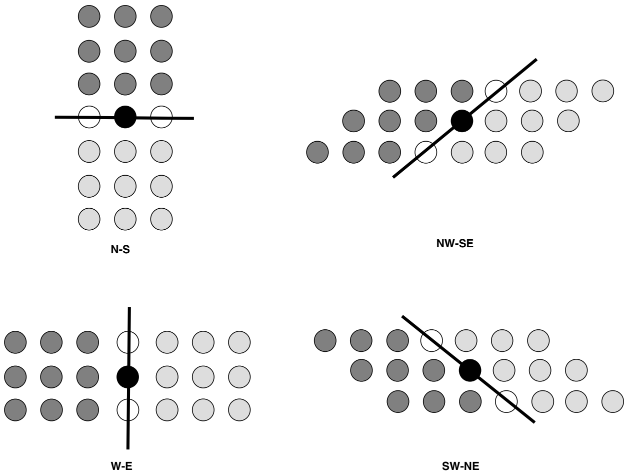

Fig. 4. Boundary information extraction. Four partitions are considered to measure the boundary probability. The maximum probability obtained is the probability to be boundary between both regions.

where α is a model parameter weighting the two terms: boundary probability and region homogeneity. R_0 – the background – is treated as a single region having a uniform probability distribution P_0 ($P_0 = 0.01$ in our experiments). It means that all pixels have a fixed probability to remain as background. A region competition algorithm [32] was applied to optimise the energy function. It takes the neighbouring pixels to the current regions boundaries $\partial\rho(R)$ into account to determine the next movement. Specifically, the optimisation process makes the most probable detected seeds (see above Section to know how to detect them) for each object to move and grow aggregating a neighbouring pixel when this new classification improves the energy. This process continues until an energy minimum is reached. At the end, the detected known objects have been segmented and classified.

2.3. Discovering unknown objects

When the minimisation process finishes, if still there is a background region R_0 which remains without being segmented/classified, it probably implies that one (or several) unknown objects are present in the image. In order to extract these objects a last stage of general purpose segmentation is performed. A new seed is placed in the background (unclassified objects), and the energy minimisation starts again. The placement of a new seed is an important choice, since in order to obtain a sample of each region large enough to statistically model its behaviour we need to place the seed completely inside one of unknown objects. A seed placed on the boundary between regions is considered as a bad seed because it would be constituted by a mixture of pixels belonging to different objects, and thus it is not adequate in order to model the region. Boundary information allows us to extract these positions in the core of regions by looking for places far away from contours. Hence, the seed is put at the position farthest away from high gradient values. Specifically, we place the seed at the place j in the background which has a lower potential defined as:

$$\text{potential}(j) = \max\left(\frac{|\nabla(i)|}{d(i,j)+1}\right)\ \forall i \in I \tag{6}$$

where $|\nabla(i)|$ is the gradient magnitude of neighbouring pixels i, $d(i,j)$ is the Euclidean distance between spatial positions of pixels i and j, and I is the image domain. One is added to the distance in order to avoid the division by zero when the influence of a pixel over itself is measured.

The seed grows guided by the optimisation and an unknown object is segmented. Note that when segmenting the region corresponding to an unknown object, all the features are used to model the region (active region segmentation) because we have not any information about the background and hidden unknown objects. This process is repeated, and a new seed placed for each unclassified object, until all the image is segmented and classified. As a result, known objects are recognised with a certain probability and unknown objects are accurately segmented.

2.4. Region belief fusion

Once the image is classified into known objects and the unknown objects are segmented, we obtain a set of disjoint regions. However, with the aim to classify unknown regions, we perform a last stage of fusion where the contextual information provided by classified neighbours is exploited. In other words, we give a higher probability to unknown regions of being classified as their neighbours (e.g. where there are bushes could be a good idea to look for more bushes). Hence, a Region Adjacency Graph (RAG) is built based on the spatial adjacency between regions [33]. Our scheme then proceeds on the RAG by defining the region belief fusion following the steps below:

(1) For all the unknown regions next to a known classified region, a similarity function (euclidean distance) using the specific features of the classified object is computed. When the result indicates a high degree of similarity ($th_r > 0.7$), both regions are merged and considered the same classified object.
(2) For all the unknown regions next to another unknown region repeat the process above. Here we use all the features to compute the similarity.
(3) Repeat steps 1 and 2 until no changes.

Fig. 5 qualitatively shows that after this last step the results are considerably improved.

In order to evaluate the proposed system, we carried out two experiments. In the first one we evaluated the performance of our system and its ability to handle with known and unknown objects. Moreover, in the second experiment we compared our technique with the results presented by Martí et al. [13] and He et al. [22] with their own datasets.

3. Data sets

We evaluated our classification algorithm on three different datasets: (i) Outex dataset [34], (ii) Martí et al. data-

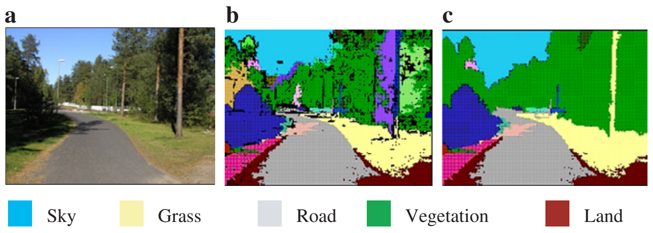

Fig. 5. Refinement of the initial classification: (a) original image; (b) initial classification; (c) refined result by exploiting the context of neighbouring regions.

set [13], and (iii) the dataset used by He et al. [22]. We will refer to these datasets as OU, MA and HE, respectively. These images consist of natural outdoor scenes and mainly contain typical objects in rural and suburban areas. Fig. 6 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. We segmented and labeled them manually into 5 classes: *sky*, *grass*, *road*, *vegetation* and *ground*, while the remaining areas, mainly belonging to man-made objects, are considered as *unknown* objects.

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 segmented and labeled them manually into the 5 same classes than in dataset OU.

HE: 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 [22]. Each image is 180 × 120 pixels. They labeled them manually 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.

4. Implementation details and methodology

4.1. Implementation details

At the learning stage, 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 components, HLS and CIE Lab* colour space [35], which is perceptually uniform. The texture information is obtained by set of co-occurrence matrix-based texture features by using a distance of one pixel and angles quantised to 45° intervals [36]. Hence, four matrices of horizontal, first diagonal, vertical, and second diagonal (0°, 45°, 90° and 135°) are used. The statistics applied were: Contrast, Homogeneity, Correlation and Entropy. Thus initially, each pixel is represented by 25 image statistics.

After carrying out the feature selection process, each object is represented by a different number of features. These features are the best which represent each object from the others. We show below the features used for each object. When working with MA dataset:

- *Sky*: G, L, S, b*, Homogeneity 0°, Homogeneity 90°, Homogeneity 145°, Correlation 45°, Correlation 90°, Entropy 45°. Its dimensionality vector is $k = 10$.

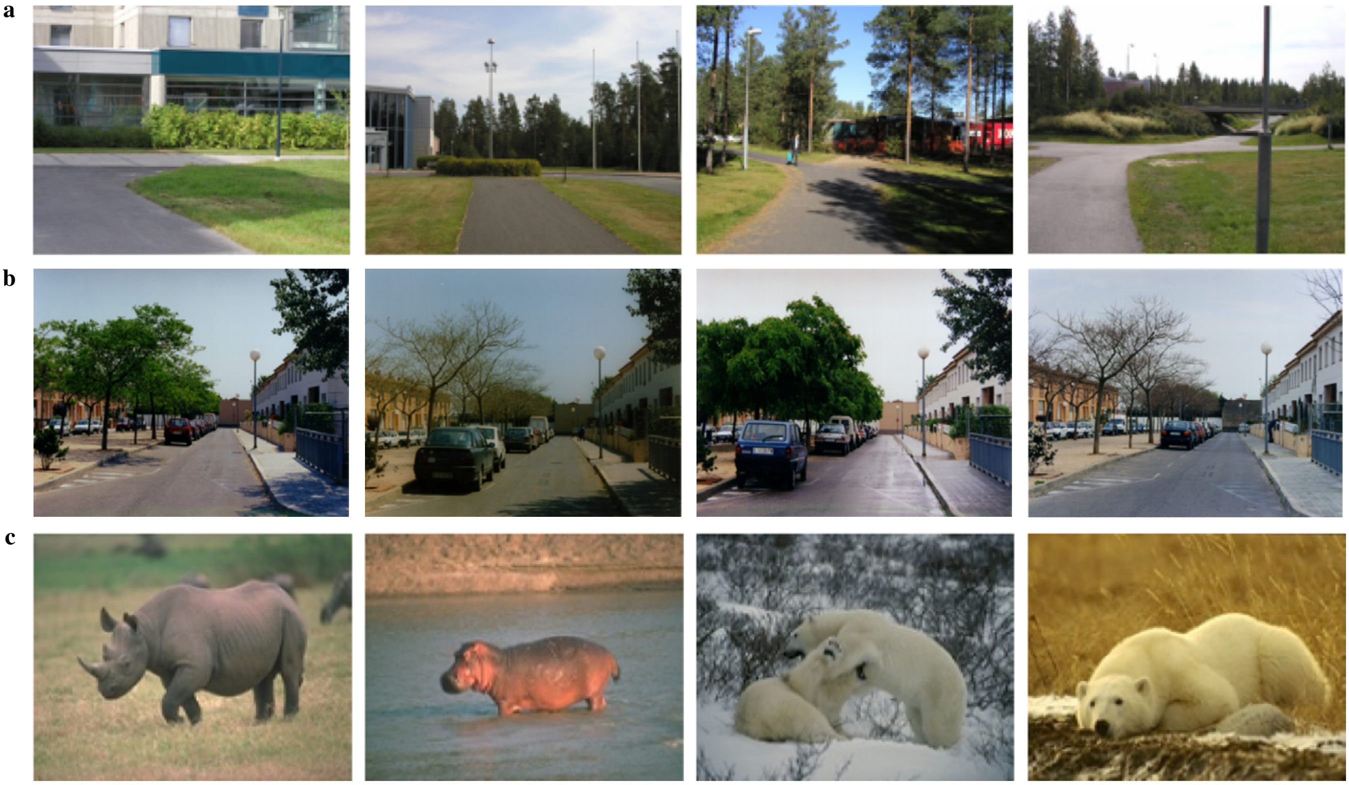

Fig. 6. Images of the datasets used to evaluate the systems: (a) images from the OU dataset [34]; (b) images from MA dataset [13]; (c) images from HE dataset [22].

- *Grass*: R, G, H, S, L*, a*, Contrast 0°, Contrast 90°, Homogeneity 45°, Correlation 45°, Entropy 45°, Entropy 90°, Entropy 145°. Its dimensionality vector is $k = 13$.
- *Road*: R, B, H, L, L*, b*, Contrast 45°, Contrast 145°, Homogeneity 45°, Homogeneity 145°, Correlation 145°, Entropy 0°. Its dimensionality vector is $k = 12$.
- *Vegetation*: G, L, a*, Contrast 90°, Homogeneity 0°, Homogeneity 90°, Correlation 0°, Correlation 90°, Entropy 0°, Entropy 45°, Entropy 90°. Its dimensionality vector is $k = 11$.
- *Ground*: R, B, L, S, b*, Contrast 0°, Contrast 145°, Homogeneity 90°, Correlation 45°, Entropy 45°, Entropy 145°. Its dimensionality vector is $k = 11$.
- *Unknown*: represented by the whole set of features. Its dimensionality vector is $k = 25$.

When working with HE dataset:

- *Rhino/hippo*: B, S, Contrast 45°, Contrast 90°, Homogeneity 45°, Homogeneity 90°, Homogeneity 145°. Its dimensionality vector is $k = 7$.
- *Polar bear*: B, L, a*, Homogeneity 45°, Homogeneity 90°, Correlation 0°, Correlation 45°. Its dimensionality vector is $k = 7$.
- *Vegetation*: G, L, L*, a*, Contrast 45°, Contrast 90°, Homogeneity 45°, Entropy 0°, Entropy 45°, Entropy 90°. Its dimensionality vector is $k = 10$.
- *Sky*: R, H, b*, Contrast 45°, Homogeneity 0°, Homogeneity 45°, Homogeneity 145°, Correlation 45°, Correlation 145°. Its dimensionality vector is $k = 9$.
- *Water*: R, G, S, Contrast 0°, Contrast 95°, Correlation 0°, Entropy 0°, Entropy 90°, Entropy 145°. Its dimensionality vector is $k = 9$.
- *Snow*: G, B, S, Contrast 45°, Homogeneity 0°, Homogeneity 90°, Correlation 90°. Its dimensionality vector is $k = 7$.
- *Ground*: R, H, L*, b*, Contrast 45°, Homogeneity 45°, Homogeneity 90°, Correlation 0°, Entropy 45°, Entropy 90°. Its dimensionality vector is $k = 10$.

Moreover to carry out with the results in this paper, we used: $P_R > 0.8$ – to accept a pixel as a seed; $\alpha = 0.7$ – the weight in formula Eq. (5); $P_0 = 0.01$ – background probability; $th_r > 0.7$ – to merge regions in region belief fusion process. The fuzzy rules are the following:

- For *top position*: if $y_j <= Y_T$ then $P_T(y_j) = 1$ otherwise $P_T(y_j)$ follows Eq. (7).
- For *middle position*: if $Y_T < y_j < Y_B$ then $P_M(y_j) = 1$ otherwise if $y_j <= Y_T$ $P_M(y_j)$ follows Eq. (9) and if $y_j >= Y_B$ $P_M(y_j)$ follows Eq. (10).
- For *bottom position*: if $y_j > = Y_B$ then $P_B(y_j) = 1$ otherwise $P_B(y_j)$ follows Eq. (8).

where P_T and P_B are:

$$P_T(y_j) = \frac{Y_{SIZE} - y_j}{Y_{SIZE} - Y_T} \tag{7}$$

$$P_B(y_j) = \frac{y_j}{Y_B} \tag{8}$$

$$P_M(y_j) = \frac{y_j}{Y_T} \tag{9}$$

$$P_M(y_j) = \frac{Y_{SIZE} - y_j}{Y_{SIZE} - Y_B} \tag{10}$$

where $Y_T = Y_{SIZE}/3$ and $Y_B = 2 * Y_{SIZE}/3$

4.2. Methodology

In order to evaluate the goodness of the implemented systems a comparison between the results of the classifications system and hand-labeled images is performed. Specifically, to know the performance of the system a confusion matrix is computed. The confusion matrix should be read as follow: columns indicate the object to recognise and rows indicate the label the system associates at this object. Hence, a perfect recognition should have 100% at all the diagonal, and zero at the remaining cells. This matrix will give us information on how the system works for each individual object. The overall performance rates are measured by the average value of the diagonal entries of the confusion matrix.

Moreover, we compared our proposal with the results obtained by a simple *pixel-based classifier*: every image pixel is classified as the object with the highest appearance probability P_A (see Eq. (1)) always this is higher tan a fixed threshold. Otherwise, the pixel is labeled as unknown. This baseline method is included in order to gauge the difficulty of the classification task. Furthermore, the improvement achieved by the inclusion of context information was quantified.

The learning stage takes approximately 40 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 1 hour and 30 minutes for a testing set of 93 images – which means about 1 minute per image (Visual C++ and php implementation on a 1.7 GHz PC).

5. Performance evaluation

We merge the two first datasets OU and MA to evaluate the performance of our proposal. We 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 (see Fig. 7).

Table 1 shows the summarised results obtained over the test image set. The pixel-based classifier achieves poor results with an accuracy of 72.64%. The inclusion of a higher region-level information by using *specific* active regions, as is proposed in our technique, allows the system

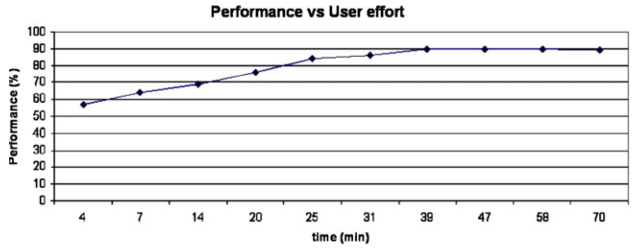

Fig. 7. Performance vs. the required time to teach the system. We start training the system with 5 images and we increment them until 50 (taking 5 images more at each next experiment).

Table 1
Quantitative results over the first test image set

Method	Pixel-based	Without ctx.	Proposal
Correct rate	72.64%	85.20%	89.87%

Correct classification rates achieved by the pixel-based classifier, the appearance-based proposal, and our whole (appearance and context) proposal.

to take the spatial consistency of objects in the image into account, improving the percentage of correctly classified pixels to 85.20%. Finally, as is shown in the last column, the conjoint use of appearance and context properties significantly improves these results and obtains a 89.87% of well-classified pixels. This shows us that the inclusion of absolute context helps to disambiguate objects with similar features.

The confusion matrix for the testing results on our proposal model is shown in Table 2. Its values show the percentage of labels on the whole testing data. We can observe that a maximum rate of 93.52% is obtained for the *sky* object, while the minimum rate is obtained for the *grass* with a 87.38%. Moreover, it is relevant to note that most of classification mistakes of our system are related to *unknown* objects, while the error between known objects is really non-frequent. This is very encouraging because this kind of errors could be solved in an easy way. Most of the pixels over-classified are confused by an unknown object (see last column in Table 2), which means not-known objects have been wrongly recognised. We consider that when the system will learn these new objects, it will be able to recognise them and the rates of over-classification will decrease considerably. On the other hand, the last row in Table 2 shows that the most important errors occur when the system does not recognise an object, and classifies it as unknown. The method is not always able to initially detect all the known objects in the image. However, since these missed objects are correctly segmented (see Fig. 8), it should be studied the possibility to correct this error by analysing the resulting unknown regions in a later stage. Exploiting the scene context in deep, we could be able to classify these objects appropriately.

Table 2
Confusion matrix over the OU and MA datasets (s, sky; g, grass; r, road; t, tree; l, land; u, unknown)

	s	g	r	t	l	u
s	93.52%	0%	0%	0%	0%	1.82%
g	0%	87.38%	0%	0%	4.17%	0.88%
r	0%	0%	91.36%	0%	4.71%	1.74%
t	0%	6.21%	0%	88.73%	0%	4.32%
l	0%	2.03%	3.39%	0%	89.97%	1.97%
u	6.48%	4.38%	5.25%	11.27%	2.15%	89.27%

Some qualitative experimental results are shown in Fig. 8. The second row shows the results achieved by our technique using only appearance properties (colour and texture), while results obtained by the whole method are shown in the third row. As previously stated, our classifier achieves a reasonable labeling of image regions. Moreover, the inclusion of context information allows to correct some mistakes performed when only the appearance was considered. In the second example (second column), the appearance-based method failed on classifying some parts of the *road* as *sky*, while the top of some trees (where leaves are confused with the sky) were wrongly recognised as *road*. All these mistakes are solved by the whole method. Furthermore, in the last stage of region fusion, the information provided by neighbouring objects also allows to correctly classify a large number of small areas of the image which were initially classified as unknown. We consider these results as very positive, although some issues need to be addressed. If we observe the last example of Fig. 8, the image classified by our proposal has a big area, corresponding to the *trees*, that is considered as an unknown region. The reason can be found in the massive presence of *shadows*, which cover this part of the image. Since we did not teach to the system to recognise the shadows, the system considers them as *unknown* objects. Therefore, we must qualify this classification as correct. This same situation can be found in the shadows on the road in the example of the first column.

5.1. Comparison

We compare our approach with the approaches of Martí et al. [13] and He et al. [22] using their own datasets MA and HE, respectively.

Martí et al. [13] proposed to classify natural objects in outdoor scenes by using binary decision trees with multivariate decision functions (a tree for each trained object class). Each node of the tree attempts to separate, in a set of known instances (the training set), a target (e.g. trees) from non-target instances (no-trees). They also use the scene context modeled as a graph with the objects that are expected to be in a scene and their spatial relationships. They used 6 texture features (Blurriness, Granularity, Discontinuity, Straightness, Curviness and Abruptness) and a set of 28 different colour features (Normalized RGB, HSV, etc.). A feature selection process is applied to find a particular subset of features for characterising each object class of interest. Their method is tested on MA and considering four natural objects (*sky*, *leaves*, *road* and *ground*). In average obtains a 87% of correctly classified pixels. This score was obtained using training

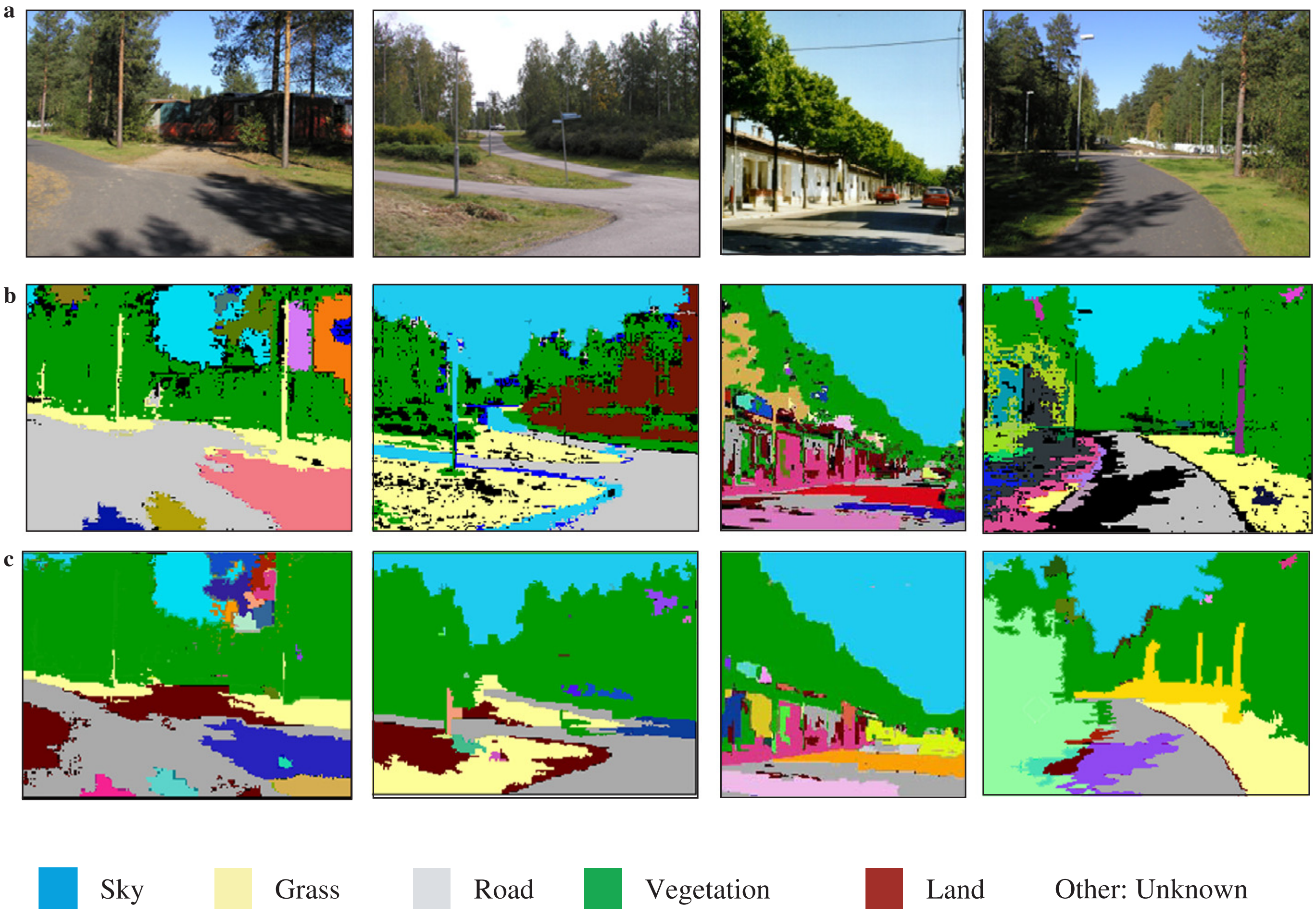

Fig. 8. Experimental results over datasets OU and MA. (a) Original image; (b) initial classification; (c) refined result by exploiting context of neighbouring regions.

and test images in the classification results evaluation. It is unusual to use the training images to evaluate the classification performance. Thus, it is very probable that if they did not include the training set the performance score would be lower. The performance is slightly improved using our approach over the same dataset. A 90.03% of well classified pixels is obtained using 20 training images and the rest (67) for testing, while classifying 5 objects (*sky*, *grass*, *ground*, *trees* and *road*). Note that we recognise 5 instead of 4 objects including *ground* as an additional object, thus the classification problem is a bit more difficult. Even so, results are improved.

The method of He et al. [22] is a multiscale Conditional Random Field (mCRF), which includes contextual features for labeling images, in which each pixel is assigned to one label of a finite set. The features are incorporated into a probabilistic framework which combines the outputs of several components. Components differ in the information they encode. Some focus on the image-label mapping, while others focus solely on patterns within the label field. Components also differ in their scale, as some focus on fine resolution patterns while others on a coarser, more global structure. A supervised version of the contrastive divergence algorithm is applied to learn these features from labeled image data. They compared their proposal with a 3-layer multilayer perceptron (MLP) and a classical Markov Random Field (MRF), and demonstrated as the inclusion of context improved considerably the results. Features used consisted on the following: for the colour information they used CIE Lab* colour space. The edge and texture properties are extracted by a set of filter banks including difference-of-Gaussian filters at 3 different scales, and quadrature pairs of oriented even- and odd-symmetric filters at 4 orientations (0, π/4, π/2, 3π/4) and 3 scales. Thus each pixel is represented by a set of 30 image statistics. In this case, the training set includes 60 randomly selected images and the remaining 40 are used for testing.

We applied our proposal over the same dataset (HE) to perform the comparison. We used 35 training images and 65 for testing. In [22] they do not take unknown objects in the images into account, so all the regions in these images are known. Similarly, we labeled all the pixels in the images without taking unknown pixels into account and we classified all the pixels of images with the most probable label. The correct classification rates on the test set are shown in Table 3. The last column in the table shows the rate obtained

Table 3
This is a comparison with the method of of He et al. [22] over the dataset D2

Method	MLP	MRF	mCRF	Proposal
Correct rate	66.9%	66.2%	80.0%	86.76%

over the same test images using our proposal. We can see that the performance of the MLP classifier is comparable to the MRF, while mCRF provides a significant improvement. The result shows the advantage of discriminative over generative modeling and the weaknesses of local interactions captured by the MRF model. However, our model increases in 6.76% the result obtained with mCRF. This could be because we use local data information of test images.

We also show the outputs of the mCRF and our model on some test images in Fig. 9. mCRF model generates reasonable labeling in which the contextual information provided by regional and global features corrects most of the wrong predictions from the local features. Also our model labels reasonably well the test images, the use of absolute context as well as data of the test image helps to disambiguate some labels and solves some of wrong labels obtained with mCRF.

5.2. More complex images

In order to evaluate the robustness of our system, we have carried out two more complex experiments: (i) evaluate the system with more complex natural scenes; and (ii) evaluate the system with rotated images. The results obtained are explained at the following:

- *Complex natural scenes*. We tested the system with more complex natural images (coast scenes with *sky* and *water* objects). We show two examples in Fig. 10. In the first row the *sky* is in almost the whole image, while in the second one the object which is in almost the whole image is the *water*. We trained the system with more simple images, where the sky is always at top and water at bottom. Fig. 10b shows the results after the *Discovering unknown objects* stage. In the first stages, where the regions grow, the location (absolute position) restricts the growing of the regions which represent the objects, and only the part of sky or water which is in its usual location is recognised. The other part of sky (or water) is classified as an unknown object. However, the region belief fusion stage, allows us to merge the unknown region and the well recognised object (sky or water in the examples) and finally the whole object is well recognised as is shown in Fig. 10c.
- *Rotated images*. We carried some experiments changing orientation of images from 10° to 180°. We can see in Fig. 11 results for the same image when it is rotated 45°, 90° and 180° grades, respectively. We observed that the results are very good if there is a part of the object which is in its habitual location (Fig. 11a and b). For

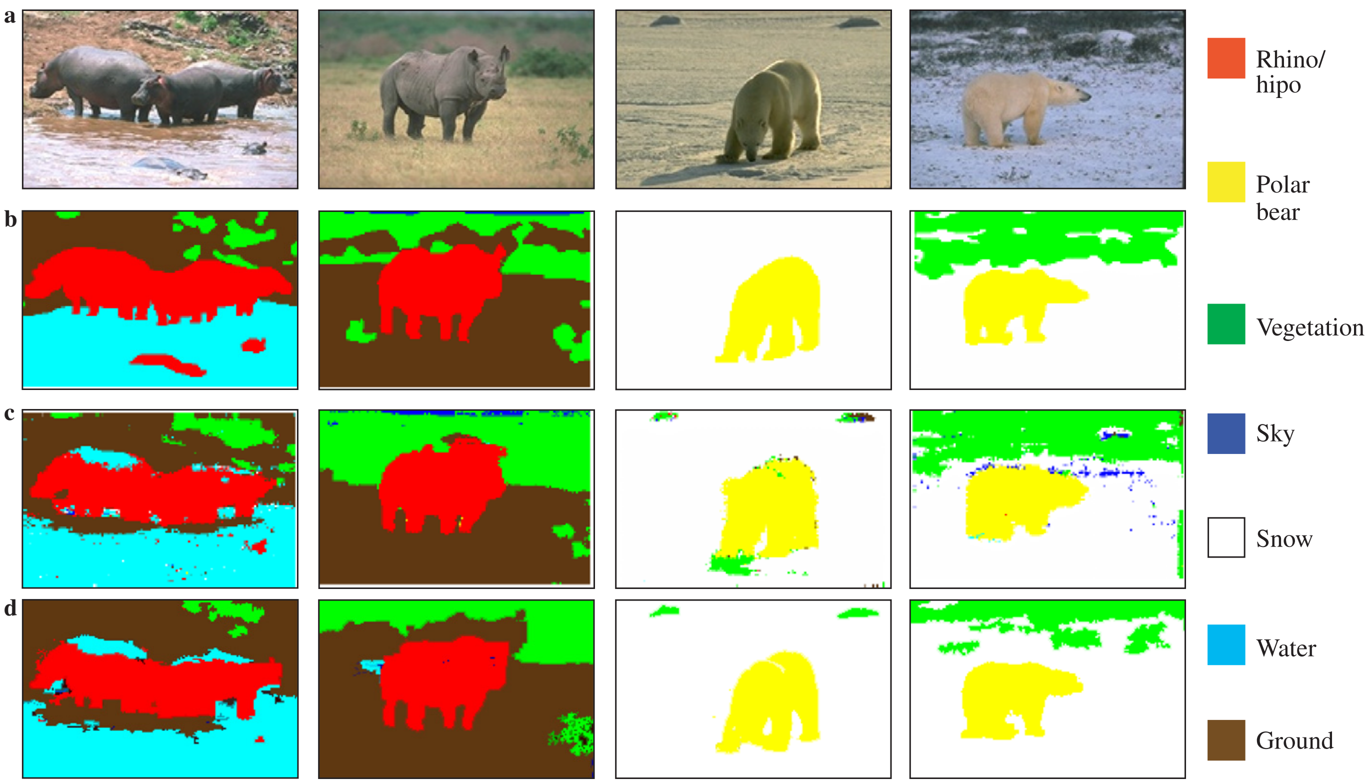

Fig. 9. Qualitative results over the Corel image dataset. (a) Original image; (b) hand-labeled image; (c) classification with mCRF; (d) classification with our proposal.

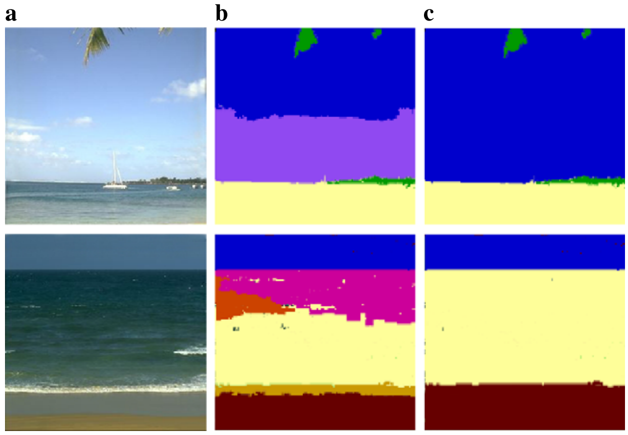

Fig. 10. Qualitative results when working with more complex natural scenes. (a) Original image; (b) image before the region belief fusion process; (c) final results (after the region belief fusion process). Blue colour represents sky, yellow means water and other colours are unknown regions.

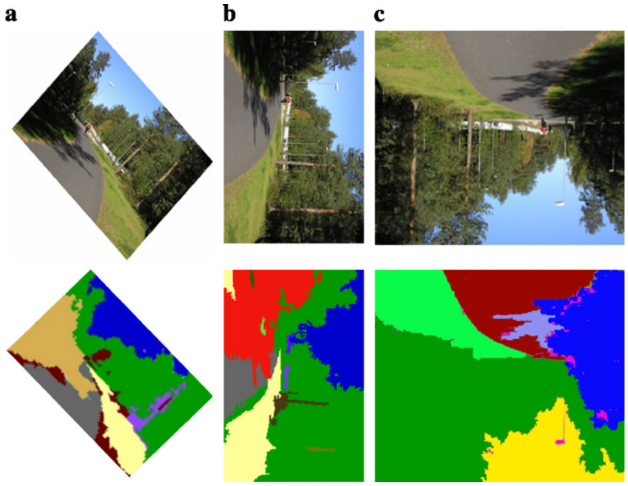

Fig. 11. Qualitative results when working with rotated images. First row shows the original image and second one shows the result. (a) Image rotated 45°; (b) image rotated 90°; (c) image rotated 180°. Colour meanings are the same as in Fig. 8.

example, if the image is rotated but there is a part of sky which is at top, then the system works well. This is because the system can place a new seed at this part of the object, and consequently recognise the part of the object which is well located. The location information prevent to recognise the part of the object which is not in its habitual position so that this one will be classified as an unknown region. However the last stage of the system, the region belief fusion, is able to merge the unknown region with the well-classified object and finally the whole object is well recognised, even if it is not in its habitual position. Nevertheless, when the image is 180° rotated the system fails when tries to recognise the objects. It can just recognise the trees, because it cannot place any other initial seed. In this case, the image is very well segmented and just the trees can be recognised.

These two experiments show the robustness of the system. We are able to deal with complex natural scenes and with rotated images. The final step which uses the neighbourhood information to merge region is the key of the system in these cases.

6. Summary and discussion

The proposed method is able to learn the model of the objects with few images. Each real object is modeled as a set of *object classes*. An *object class* is a prototype of a real object described in terms of colour and textural features under specific outdoor conditions like in [13]. At the classification stage, we take advantage of the information provided for the test image. In [18] they constraint the object model using the distribution of the newly observed data, while in our approach we do the reverse: we extend the object model using the newly test data distribution. Moreover in [21], they used active regions segmentation to obtain the regions of the image. We extended this work using *specific* active regions taking advantage of the knowledge to use the best features to classify each specific object. So we are able to recognise the learned objects and segment the unknown.

In this framework, the detection and recognition of objects proceed simultaneously with image segmentation in a competitive and cooperative manner. The method makes use of bottom-up proposals combined with top-down generative models. Tu et al. [37,38] presented an approach with a similar philosophy but it is only applied to classify and distinguish text and faces, while our work is applied with a great variety of images in different and more difficult variable conditions. Moreover we tackle the unknown objects. In many systems, the problem of unknown objects is engineered away or resolved in a preprocessing step [39], while the proposed system is able to tackle this objects in an *on-line* manner for a further knowledge of them.

Our first experiment consisted on evaluating our method using natural outdoor scenes, mainly containing typical objects in rural and suburban area (datasets OU and MA). First, we evaluated if the inclusion of test data in the learned model and the inclusion of higher region-level information improves the obtained results when classifying the images. We obtain an improvement of 12% approximately. Then we evaluated how important is the use of absolute context and local relationships between objects as well. Using both, we obtain an score of 89.87% that is a 4.67% more than without using context information. The use of this kind of context helps to disambiguate some mistakes occurred, for example the when system confuses some parts of the *sky* as *road*.

The comparison experiment consisted on comparing the performance of our system to other Image Understanding and object classifications systems. We compare the model with the systems proposed in [13,22]. In the first work, the classification is carried out using a decision tree including scene and relative context. In this work, the input to the system is the test image and its kind of scene. In the second work, the proposed model, a mCRF, was compared with a simple MLP and a MRF. Our proposal has been compared with all previous methods. In order to make this comparison, we applied our method to the same images as them and evaluated the results using the same ground-truth. These results show that our system obtains a 3% in first case and a 6.76% in the second comparison better classification rates. This could be due to the use of test data at the classification stage and to the use of local relationships between objects in order to improve the classification of final unknown objects.

Moreover we showed that the system is able to work with more complex natural scenes and with rotated images even thought we are using absolute contextual information. As we showed in Section 5.2, the system is able to deal with these difficult situations because we are using a last stage of *region belief fusion*. However for 180° rotated images, the system can recognise few objects, but the segmentation is very good. Although this kind of situation is not usual in most of the natural scenes databases, it could be solved with a pre-processing step to detect the image orientation.

By the moment, we do not include the context provided by the scene configuration. The reason is that the scene model is referred to a certain combination of objects, for example if we have a zoo scene, then we will expect to find a rhino, an hippo and other animals. Thus, the knowledge of the scene type can help and make the recognition easier. On the other hand, the use of rules and constraints too much strict can avoid the system recognises images and objects that differ of these models, while we humans are able to recognise an hippo in the middle of the city. Thus, we consider the inclusion of this information must be carefully designed to guarantee the system is able to handle with unknown situations.

7. Conclusions and further work

We have presented a probabilistic model for labeling images into a set of learned class labels, and segmenting the unknown objects. The model combines the data acquired during the learning stage as well as the data of the actual test image in order to obtain a more accurate result. Moreover, the labels are in agreement with the image statistics and with the absolute contextual information as well. The object extraction and recognition is carried out by the integration of an initial pixel-level classification, which provides the CO, and a later growing of *specific* active regions, which allows to take the spatial consistency of objects into account. This growing was done by optimising an energy function using the region competition algorithm.

We have presented the results of our probabilistic model for labeling images into a predefined set of class labels. The results show that it is useful not only the use of the models acquired during the learning, but also the use of test data to improve them. Our strategy results in a consensual labeling that needs to agree with the image statistics and at the same time respect the absolute position in the image. Moreover, we compared our system with two recent systems proposed in 2001 and 2004 and results we obtained are superior. We tested the method using natural outdoor scenes.

In the future we would like to test this method with man-made and indoor images. And also using more recent features invariant to geometric, photometric (e.g. features proposed in [40]) and scale (e.g. SIFT features [41]) of the objects. We also will improve the approach given the possibility to learn the unknown segmented objects obtained after the classification. Techniques for an unsupervised learning and recognition of them will be explored.

Acknowledgements

We would like to thank X. He, R.S. Zemel and M.Á Carreira-Perpiñán at the University of Toronto, for providing their image data and corresponding ground-truth. This work was partially founded by research Grant BR03/01 from the University of Girona.

References

[1] A. Vailaya, A. Figueiredo, A. Jain, H. Zhang, Image classification for content-based indexing, IEEE Transactions on Image Processing 10 (2001) 117–130.

[2] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, D. Forsyth, The effects of segmentation and feature choice in a translation model of object recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, Springer, Madison, Wisconsin, 2003, pp. 675–682.

[3] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei, M. Jordan, Modeling words and pictures, Machine Learning (3) (2003) 1107–1135.

[4] C. Pantofaru, R. Unnikrishnan, M. Hebert, Toward generating labeled maps from color and range data for robot navigation, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 2, Las Vegas, Nevada, 2003, 1314–1321.

[5] F. Li, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA, 2005, pp. 524–531.

[6] 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.

[7] J. Sivic, B.C. Russell, A. Efros, A. Zisserman, W.T. Freeman, Discovering object categories in image collections, Tech. Rep. AI Memo 2005-005, Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, 2005.

[8] A. Torralba, Contextual priming for object detection, International Journal of Computer Vision 53 (2) (2003) 169–191.

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

[10] A. Singhal, J. Luo, W. Zhu, Probabilistic spatial context models for scene content understanding, in: IEEE Computer Society Conference

on Computer Vision and Pattern Recognition, vol. 1, Madison, Wisconsin, 2003, pp. 235–241.

[11] E. Sudderth, A. Torralba, W. Freeman, A. Willsky, Learning hierarchical models of scenes, objects and parts, in: International Conference on Computer Vision, vol. 2, Beijing, China, 2005, pp. 1331–1338.

[12] J. Batlle, A. Casals, J. Freixenet, J. Martí, A review on strategies for recognizing natural objects in colour images of outdoor scenes, Image and Vision Computing 18 (2000) 515–530.

[13] J. Martí, J. Freixenet, J. Batlle, A. Casals, A new approach to outdoor scene description based on learning and top-down segmentation, Image and Vision Computing 19 (2001) 1041–1055.

[14] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, International Journal of Computer Vision 42(3) 145–175.

[15] M. Hudelot, M. Thonnat, A cognitive vision platform for automatic recognition of natural complex objects, in: IEEE International Conference on Tools with Artificial Intelligence, Sacramento, California, USA, 2003.

[16] C. Chu, A. Aggarwal, Image interpretation using multiple sensing modalities, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (8) (1992) 840–847.

[17] V.P. Kumar, U.B. Desai, Image interpretation using bayes networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (1996) 74–77.

[18] S. Kumar, A.C. Loui, M. Hebert, An observation-constrained generative approach for probabilistic classification of image regions, Image and Vision Computing 21 (2003) 87–97.

[19] T. Drummond, Learning task-specific object recognition and scene understanding, Computer Vision and Image Understanding 80 (2000) 315–348.

[20] T. Strat, M. Fischler, Context-based vision: Recognizing objects using information from both 2d and 3d imagery, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 1050–1065.

[21] J. Freixenet, X. Muñoz, J. Martí, X. Lladó, Color texture segmentation by region-boundary cooperation, in: European Conference on Computer Vision, Vol. II, Prague, Czech Republic, 2004, pp. 250–261.

[22] X. He, R.S. Zemel, M. Á. Carreira-Perpiñán, Multiscale conditional random fields for image labeling, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, Washington DC, USA, 2004, pp. 695–702.

[23] N. Serrano, A. Savakis, J. Luo, Improved scene classification using efficient low-level features and semantic cues, Pattern Recognition 37 (2004) 1773–1784.

[24] J. Fan, Y. Gao, H. Luo, G. Xu, Statistical modeling and conceptualization of natural images, Pattern Recognition 38 (2005) 865–885.

[25] P. Viola, M. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.

[26] F. Li, R. Fergus, P. Perona, A bayesian approach to unsupervised one-shot learning of object categories, in: International Conference on Computer Vision, vol. 2, Nice, France, 2003, pp. 1134–1141.

[27] F. Li, R. Fergus, P. Perona, Learning generative visual models from few training examples, in: Workshop on Generative-Model Based Vision, 2004.

[28] R. Fergus, F. Li, P. Perona, A. Zisserman, Learning object categories from google's image search, in: International Conference on Computer Vision, Beijing, China, 2005.

[29] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (11) (1994) 1119–1125.

[30] K. Barnard, D. Forsyth, Learning the semantics of words and pictures, in: International Conference on Computer Vision, vol. 2, 2001, pp. 408–415.

[31] N. Paragios, R. Deriche, Geodesic active regions and level set methods for supervised texture segmentation, International Journal of Computer Vision 46 (3) (2002) 223–247.

[32] S. Zhu, A. Yuille, Region competition: unifying snakes, region growing, and bayes/mdl for multi-band image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (9) (1996) 884–900.

[33] J. Luo, C. Guo, Perceptual grouping of segmented regions in color images, Pattern Recognition 36 (2003) 2781–2792.

[34] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, S. Huovinen, Outex – new framework for empirical evaluation of texture analysis algorithms, in: IAPR International Conference on Pattern Recognition, vol. 1, Québec City, 2002, pp. 701–706.

[35] M. Celenk, A color clustering technique for image segmentation, Computer Vision, Graphics and Image Processing 52 (1990) 145–170.

[36] R. Haralick, K. Shanmugan, I. Dunstein, Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics 3 (1973) 610–621.

[37] Z. Tu, X. Chen, A.L. Yuille, S. Zhu, Image parsing: Unifying segmentation, detection, and recognition, in: International Conference on Computer Vision, vol. 1, Nice, France, 2003, pp. 18–25.

[38] Z. Tu, X. Chen, A.L. Yuille, S. Zhu, Image parsing: unifying segmentation, detection, and recognition, International Journal of Computer Vision 63 (2) (2005) 113–140.

[39] B. Milch, B. Marthi, S. Russell, D. Sontag, D.L. Ong, A. Kolobov, Blog: Probabilistic models with unknown objects, in: International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 2005, pp. 1352–1359.

[40] F. Schaffalitzky, A. Zisserman, Viewpoint invariant texture matching and wide baseline stereo, in: International Conference on Computer Vision, Vol. 2, Bancouver, B.C., Canada, 2001, pp. 636–643.

[41] D. Lowe, Distinctive image features from scale invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.