

Classifying Natural Objects on Outdoor Scenes

Anna Bosch ¹, Xavier Muñoz, Joan Martí, Arnau Oliver

*University of Girona
Computer Vision and Robotics Group*

Abstract. We propose an hybrid and probabilistic classification of image regions belonging to scenes primarily containing natural objects, e.g. sky, trees, etc. as a first step in solving the problem of scene context generation. Therefore, we will focus our work in the problem of image regions labeling to classify every pixel of a given image into one of several predefined classes. Our proposal begins with a top-down control to find the core of objects, which allow us to update the learned models. Moreover, they become the starting seeds for the growing of a set of concurrent active regions which, considering the own region model as well as region and boundary information, obtain an accurate recognition of known regions. Next, a general segmentation extracts the unknown regions by a bottom-up strategy. Finally, a last stage exploits the contextual information to classify initially unknown segmented objects. 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. Experimental results on a wide set of outdoor scene images are shown to evaluate and compare our proposal.

Keywords. Computer Vision, Object Classification, Image Understanding

1. Introduction

Modern research in image understanding has naturally evolved from a prior research in image analysis. Part of this evolution involves moving from low level image description to more meaningful semantic interpretation [1]. Automatically labeled images and automatic extraction of the semantic context of a scene can also be useful for other purposes, such as for image indexing and retrieval, robotic navigation, surveillance, object detection and recognition.

We tackle in this paper the problem of image labeling: to classify every pixel of a given image into one of several predefined classes. Moreover, we focus on the recognition of natural objects. Hence, we might consider images of outdoor scenes and we would like to classify each pixel as sky, grass, trees, etc. To achieve this goal, and in 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, [2] used

¹Correspondence to: Campus de Montilivi s/n. 17071 Girona, Spain Tel.: 971418486; Fax: 972418976; E-mail: aboschr@eia.udg.es

texture features in order to classify textured surfaces, such as sky, forest, ground and sea, in outdoor images, while [3] considered colour, texture and shape information to generate maps segmented into objects of interest: buildings, vegetation, etc. 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 [4]. The main drawback of not using context is the overlap between classes, e.g. sky and water, both blues. The system can then easily confuse a water region, at the bottom of the image, with the sky, since they have a very similar appearance. Two small image patches are ambiguous at a very local scale but clearly identifiable inside their context.

The proposed method is an hybrid and probabilistic classification (taking appearance and contextual information into account) of image regions belonging to scenes primarily containing natural objects, as a first step in solving the problem of scene context generation. Furthermore, the technique handles with known and unknown objects in the image by following an hybrid control: it begins with a top-down control based on specific active regions to find the known objects, and following a general segmentation provides the segmentation of unknown regions by a bottom-up strategy. Finally, the use of contextual information given by the neighbouring regions allows us to refine the initial classification of unknown objects. 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.

This paper is organised as follows. Section 2 describes our proposal, taking the phase of recognition into account, and specially focusing on the detection of known objects core and the later growing of a concurrent set of active regions. In Section 3 some experimental results are shown and discussed. We evaluate the performance of our system and its ability to handle with known and unknown objects. Moreover, the results are compared with a relevant and recent work of 2004 [5]. 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 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 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 of the remaining ones (see [6] for further details of the learning stage). Next, our proposal starts the recognition by using the knowledge of the learned objects to obtain the probability of each pixel to belong to each object, which provides us the probabilistic pixel maps (one map for each object). The main contribution in our approach lies in the next stage: the most probable pixels of each map are detected, which constitute the core of objects, and are used to extract a new and more accurate object model. We consider the core pixels as samples of the regions and the object model is updated to match with the regions present in the image. The posterior growing of specific active regions from these cores allows to classify and segment the image. Until here the algorithm follows a top-down con-

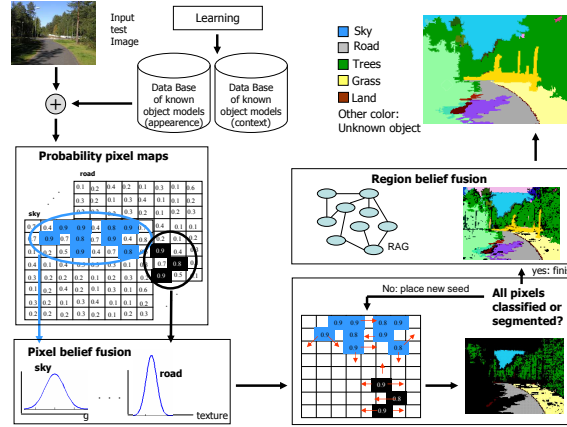


Figure 1. Proposed hybrid method for the classification and segmentation of the image.

trol, since the knowledge is used at the beginning of the process. Following, a bottom-up control is applied by performing a general purpose segmentation of not-classified 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. Figure 1 shows the basic architecture of our proposal.

2.1. 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 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 [3].

2.1.1. 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 \vec{x}_j of belonging to a object \mathcal{O}_{Li} is given, under a Gaussian assumption, by the probability density function:

$$P_A(j|\mathcal{O}_{Li}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma_i|}} \exp\left\{-\frac{1}{2}(\vec{x}_j - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x}_j - \vec{\mu}_i)\right\} \quad (1)$$

where $\vec{\mu}_i$ is the mean vector of the object \mathcal{O}_{Li} , Σ_i its covariance matrix, and k the number of characteristics.

At this stage, we compute a contextual probability by using a fuzzy rule based approach. 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 object is computed 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, equation 2 gives us the probability a pixel j at position y_j belongs to an object \mathcal{O}_{L_i} considering its absolute position:

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

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 avoiding its confusion. Therefore, the merging of both probabilities (P_R) provides a probabilistic pixel map for each object.

2.1.2. Pixel belief fusion

Nevertheless, there are only a few pixels with a very high probability to belong to a certain object. In other words, a reduced set of pixels can be classified at this time, with a high confidence of being taking the right decision. This is due to the fact that few images have been used in the learning, 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 given one we are trying to recognise.

Similarly to the proposal of [7] 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 constitute the object core, and are considered as representative data to design a lesser constrained new model. For each object, $\vec{\mu}_i$ and Σ_i , which characterises the model, are re-computed in order the model represents the reality of the test image. This new set of objects is called \mathcal{O}_N : $\mathcal{O}_N = [\mathcal{O}_{N1}(\vec{\mu}_1, \Sigma_1), \dots, \mathcal{O}_{Nn}(\vec{\mu}_n, \Sigma_n)]$.

2.1.3. Object belief refinement

The core pixels are used as starting seeds to initialise the growing of a concurrent set of 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.

Moreover, 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. The region 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 tougher 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 of the pixel (referred as m and its opposite as n). Therefore, $P_R(m|A)$ is the probability that features obtained in the side m belong

to object A , while $P_R(n|B)$ 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 $P_R(m|A) \times P_R(n|B)$, which is maximum when j is exactly the edge between objects A and B because both sides fit better with both models. Four possible neighbourhood partitions (vertical, horizontal and two diagonals) are considered as in the proposal of [8]. Therefore, the corresponding probability of a pixel j to be boundary, $P_B(j)$, is the maximum probability obtained on the four possible partitions.

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

$$E(\rho(R)) = (1 - \alpha) \sum_{i=1}^N -\log P_B(j : j \in \partial R_i) + \alpha \sum_{i=0}^N -\log P_R(j : j \in R_i | R_i) \quad (3)$$

where α is a model parameter weighting the two terms: boundary probability and region homogeneity. A region competition algorithm [9] was applied to optimise the energy function. Intuitively, all regions begin to move and grow, competing for the pixels of the image until an energy minimum is reached. At the end, the detected known objects have been segmented and classified.

2.2. Discovering unknown objects

If still there are areas of the image which remain 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, and the energy minimisation starts again. Note that when segmenting the region corresponding to an unknown object, all the features are used to model the region. This process is repeated, and a new seed launched for each not-classified object, until all the image is segmented. As result, known objects are recognised with a certain probability and unknown objects are accurately segmented.

2.2.1. 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 looking for more bushes). Hence, a Region Adjacency Graph (RAG) is built based on the spatial adjacency between regions. Our scheme then proceeds on the RAG by defining the region belief fusion. If an unknown region is near a known classified region, a similarity function is computed. When the result indicates a high degree of similarity, both regions are merged and considered the same object.

3. Experimental results

3.1. Data Sets

We applied our method to a colour image data set constructed using 125 images from the Outex image database [10], and also a set of images taken by ourselves. These images consist of natural outdoor scenes and mainly contain typical objects in rural and suburban areas. We segmented and labeled them manually into 5 classes: *sky*, *grass*, *road*, *vegetation* and *land*, while the remaining areas, mainly belonging to man-made objects, are considered as *unknown* objects.

The second image dataset 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 [5]. 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.

3.2. Learning

Each training set includes 35 selected images and the remaining ones for testing. This number of training images was stated in our experiments as a good compromise between the required effort of the user and the quality of results. For these experimental trials, a large number of colour and texture features were initially considered as candidates to be selected to describe the objects: RGB, HLS and CIE Lab* colour space, and a set of 8 co-occurrence matrix-based texture features. The learning stage takes approximately 30 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 1 hour and 30 minutes for each set of testing images (Visual C++ and php implementation on a 1.7GHz PC).

3.3. Evaluation and Comparison

The method achieved a percentage of 89.87% of well-classified pixels over the Outex dataset. Moreover, the confusion matrix for the testing results is shown in Table 1. We must note that most of classification mistakes are related to *unknown* objects, while the error between known objects is really non-frequent. The system sometimes wrongly labels an unknown object, or contrarily it misses a known object. Fortunately, both kind of errors could have a probable solution a) when the system learns these new objects, and b) analysing the resulting unknown regions in a later stage with the aim of trying to classify unknown objects by exploiting the object model as well as the scene context information. Some qualitative experimental results are shown in Figure 2. Note that the known objects are correctly classified, while unknown areas are accurately segmented, although some issues need to be addressed. If we observe the last column of Figure 2, a big area, corresponding to the trees, has been 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 the system to recognise the shadows, the system considers them as *unknown* objects. Therefore, we must qualify this classification as correct.

We also applied our proposal over the Corel dataset to perform a comparison with the results published in 2004 by He et al. [5]. The method of He et al. is a multiscale Con-

	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%

Table 1. Confusion matrix over the Outex dataset (s=sky, g=grass, r=road, t=tree, l=land, u=unknown).

	r/h	pb	w	sn	v	g	sk
r/h	85.52%	5.25%	5.34%	2.53%	1.55%	4.63%	0%
pb	2.12%	81.38%	0.1%	4.52%	2.40%	0.32%	0%
w	5.21%	0%	89.62%	0%	1.12%	1.77%	3.04%
sn	0%	8.76%	0%	83.96%	3.45%	0.56%	0%
v	0%	0%	1.66%	8.99%	88.47%	3.45%	7.84%
g	7.15%	4.61%	3.28%	0%	1.45%	89.27%	0%
sk	0%	0%	0%	0%	1.56%	0%	89.12%

Table 2. Confusion matrix over the Corel dataset (r/h=rhino/hippo, pb=polar bear, w=water, sn=snow, v=vegetation, g=ground, sk=sky).

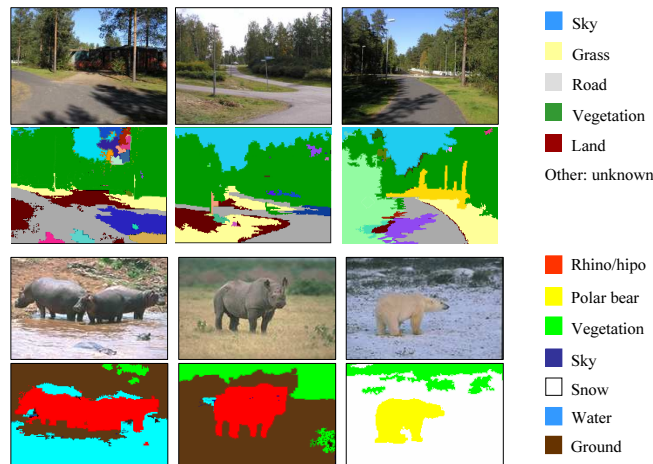


Figure 2. Some experimental results for the Outex (2 top rows) and Corel (2 bottom rows) datasets.

ditional Random Field (mCRF), which includes contextual features at different scales for labeling images. They compared their proposal with a 3-layer multilayer perceptron and a classical Markov Random Field, and demonstrated as the inclusion of context improved considerably the results. The classification rate obtained by our technique was of 86.76%, which is very encouraging because it improves in 6.76% the result obtained with the mCRF. Moreover, Table 2 shows the confusion matrix for the testing results, which proves that our technique is consistent across the classes.

4. 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 core of objects, 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. In the future we will try to test more recent methods to optimise it.

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] J. Luo and P. Etz Stephen, "Improved scene classification using efficient low-level features and semantic cues," *Pattern Recognition*, vol. 37, pp. 1773–1784, September 2002.
- [2] D. Puig and M. A. Garcia, "Pixel classification through divergence-based integration of texture methods with conflict resolution," in *IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003, vol. II, pp. 1037–1040.
- [3] C. Pantofaru, R. Unnikrishnan, and M. Hebert, "Toward generating labeled maps from color and range data for robot navigation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Las Vegas, Nevada, October 2003, vol. 2, pp. 1314–1321.
- [4] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.
- [5] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán, "Multiscale conditional random fields for image labeling," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington D.C., USA, June 2004, vol. 2, pp. 695–702.
- [6] X. Muñoz, A. Bosch, J. Martí, and J. Espunya, "A learning framework for object recognition on image understanding," in *Iberian Conference on Pattern Recognition and Image Analysis*, Estoril, Portugal, June 2005.
- [7] S. Kumar, A. C. Loui, and M. Hebert, "An observation-constrained generative approach for probabilistic classification of image regions," *Image and Vision Computing*, vol. 21, pp. 87–97, 2003.
- [8] N. Paragios and R. Deriche, "Geodesic active regions and level set methods for supervised texture segmentation," *International Journal of Computer Vision*, vol. 46, no. 3, pp. 223–247, 2002.
- [9] S.C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing, and bayes/mdl for multi-band image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 884–900, September 1996.
- [10] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, and S. Huovinen, "Outex - new framework for empirical evaluation of texture analysis algorithms," in *IAPR International Conference on Pattern Recognition*, Québec City, August 2002, vol. 1, pp. 701–706.