USING APPEARANCE AND CONTEXT FOR OUTDOOR SCENE OBJECT CLASSIFICATION

A. Bosch, X. Muñoz and J.Martí

University of Girona
Computer Vision and Robotics Group
Campus de Montilivi s/n. 17071 Girona, Spain

ABSTRACT

We propose a probabilistic object classifier for outdoor scene analysis as a first step in solving the problem of scene context generation. The method begins with a top-down control, which uses the previously learned models (appearance and absolute location) to obtain an initial pixel-level classification. This information provides us the core of objects, which is used to acquire a more accurate object model. Therefore, their growing by specific active regions allows us to obtain an accurate recognition of known regions. Next, a stage of general segmentation provides the segmentation of unknown regions by a bottom-strategy. Finally, the last stage tries to perform a region fusion of known and 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. Furthermore, experimental results are shown and evaluated to prove the validity of our proposal.

1. INTRODUCTION

In absence of any prior information, the scene classification task requires previous knowledge about objects contained in the image. There are a lot of researchers that assume as knowledge only the appearance of objects (color, texture and shape). As recent examples, Puig and García [1] used texture features in order to classify textured surfaces, such as sky, forest, ground and sea, in outdoor images. Pantofaru et al. [2] considered color, texture and shape information to generate maps segmented into objects of interest, which are labelled according to its type: buildings, vegetation, etc.

Furthermore, it is increasingly being recognized in the vision community that context information is necessary for reliable extraction of the image regions and objects. Experiments in scene perception and visual search, have shown that the human visual system makes extensive use of this contextual information for facilitating object detection and recognition in the early stages of the recognition process [3]. 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. Specifically, we distinguish two kinds of context information: (i) *Absolute context*: refereed to the location of objects in the image (sky is at top of the image, and water at bottom), (ii) *Relative context*: position of the objects respect to other objects in the images (grass tends to be next to the road, and clouds in the sky). Some proposals consider both kinds of context [4], while only the relative context is considered by He et al. [5].

Our goal is to develop a probabilistic object classifier, which is mainly based on a probability density function (taking appearance and absolute context into account), and a posterior object-specific active region segmentation. Next, the contextual information given by the adjacency of 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 organized as follows. Section 2 describes our proposal, focusing on the phase of recognition. In Section 3 we introduce the method used to test our experiments and discuss the results on five real-world categories of different objects. We finish the paper with the conclusions and further work.

2. SYSTEM OVERVIEW

Three questions have to be addressed in order to pursue our idea: How to obtain the classification and segmentation of the known and unknown objects of the test image? How to use contextual information? Which control strategy must be the best one to obtain our goals? In this Section we address these questions in a Fuzzy and Bayesian setting and by an specific active region-based segmentation.

We propose to solve these questions by using few im-

This work was partially founded by research grant BR03/01 from the University of Girona

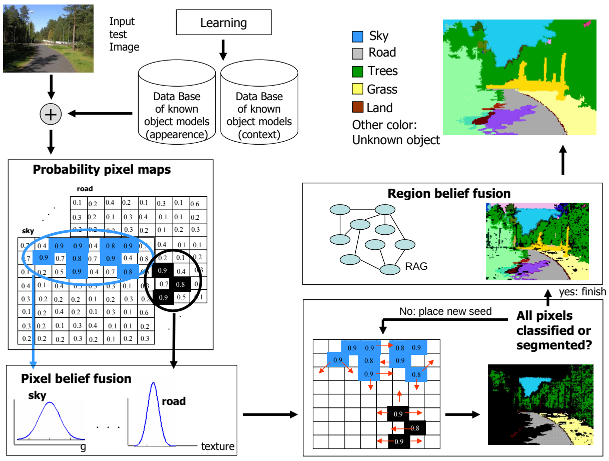

Fig. 1. Proposed hybrid method for the classification and segmentation of the image.

ages 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 differentiate the current object to the remaining ones (see [6] for more details of the learning stage). Next, our proposal starts the recognition by using the knowledge of the learned objects to obtain the probability of every pixel of belonging 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 as samples to extract a new and more accurate model that uses as object characteristics the information given by the pixels of the current test 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 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 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 neighboring objects to refine the initial classification of unknown regions. Figure 1 shows the basic schema of our proposal's architecture.

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.

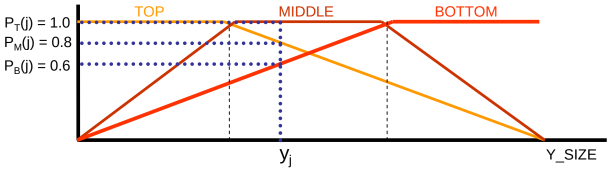

Fig. 2. Fuzzy rules for the initial context information, which provide the position of a pixel in the image. The origin 0 of Y_Size is considered at the top of the image.

The appearance probability of a pixel j characterized by the features $\vec{x_j}$ of belonging to a object $\varnothing_i$ is given, under a gaussian assumption, by the probability density function:

$$P_A(j|\varnothing_i) = \frac{1}{\sqrt{(2\pi)^k|\Sigma_i|}}\exp\{-\frac{1}{2}(\vec{x_j}-\vec{\mu_i})^T\Sigma_i^{-1}(\vec{x_j}-\vec{\mu_i})\} \tag{1}$$

where $\vec{\mu_i}$ is the mean vector of the object i, 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. Figure 2 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, equation 2 gives us the probability a pixel j at position y_j belongs to an object $\varnothing_i$ considering its absolute position:

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

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 avoid its confusion. Therefore, the merging of both probabilities allows to obtain a probabilistic pixel map for each object.

2.2. Pixel belief fusion

Nevertheless, there are only a few pixels with a very high probability to belong to a certain object and, that can be classified at this time with a high confidence of being taking the right decision. 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 recognize. 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 less constrained new model. For each object, $\overrightarrow{\mu_i}$ and Σ_i, which characterizes the model, are re-computed in order the model represents the reality of the test image. This new set of objects is called $\emptyset_N$: $\emptyset_N$=[$\emptyset_{N1}(\overrightarrow{\mu_1}, \Sigma_1)$,...,$\emptyset_{Nk}(\overrightarrow{\mu_k}, \Sigma_k)$].

2.3. Object classification and segmentation

The core pixels are then used as starting seeds to initialize the growing of a set of active regions. In [7], we presented our previous proposal of active region segmentation integrating region and boundary information, which was initially applied to unsupervised color texture segmentation. Here, the technique is extended to the problem of object recognition. Regions start to grow from the core pixels guided by the specific object model as the image data in order to segment the whole object. 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.

However, at the end of this process, if still there are areas of the image which remain without being 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 minimization starts again looking for a new optimal classification. This process is repeated until all the image is segmented. As result, known objects are recognized 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 neighbors is exploited. In other words, we give a higher probability to unknown regions of being classified as their neighbors (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. Figure 3 shows by a qualitatively way that after this last step the results are considerably improved.

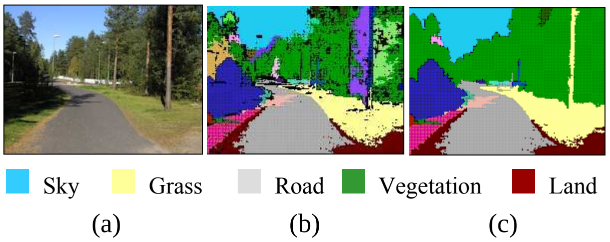

Fig. 3. Refinement of the initial classification . (a) Original image, (b) initial classification, (c) refined result by exploiting context of neighboring regions.

3. EXPERIMENTAL RESULTS

We applied our method to a set 125 color images from the image database of the University of Oulu [8], and also a set of images taken by ourselves. These images consists on natural outdoor scenes and mainly contain typical objects in rural and suburban area. We segmented and labelled them manually into 5 classes: *sky*, *grass*, *vegetation*, *road*, and *land*. The remaining areas are considered as *unknown* objects. The training set includes 35 selected images and the remaining 90 for testing. This number of training images was stated in our experiments as a good compromise between the required user effortless and the quality of results. For the experimental trials shown in this paper, a large number of color and texture features were initially considered as candidates to be selected to describe the objects: 28 color features related to different color spaces, and a set of 8 co-occurrence matrix-based texture features. The system is available on an on-line web-based application at *http://ryu.udg.es/indexant.php*.

In order to evaluate the performance of our classification method, the percentage of correctly classified pixels and wrongly over-classified pixels were measured. 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 Section 2.1), whenever this is higher tan a fixed threshold. Otherwise, the pixel is labelled as unknown. Furthermore, the improvement achieved by the inclusion of context information was quantified. Results obtained by our technique using only appearance properties and the whole method were evaluated.

Table 1 shows the summarized results obtained over the test image set. The pixel-based classifier achieves a poor results with an accuracy of 54.21%; while the inclusion of a higher region-level information by using specific active regions, as is proposed in our technique, allows the system to take the spatial consistency of objects in the image into account, which improves the percentage of correctly classified pixels to 85.20%. Finally, as is shown in the last col-

Evaluation	Pixel-Based		Without ctx.		Proposal	
	Avg	Std	Avg	Std	Avg	Std
Classified	54.21%	8.15%	85.20%	4.65%	89.87%	2.20%
Over-classified	3.05%	3.32%	2.22%	1.82%	0.90%	0.89%

Table 1. Quantitative results over the test image set. Correct classification and over-classification rates achieved by the pixel-based classifier, the appearance-based proposal, and our whole (appearance and context) proposal.

umn, the conjoint use of appearance and context properties significantly improves these results and obtains a 89.87% of well-classified pixels. Moreover, if we focus our attention on the percentage of over-classified pixels, the percentage of error also decreases in our proposal.

Some experimental results achieved by our technique are shown in Figure 3.c. As is stated, our classifier achieves a reasonable labelling of image regions. Moreover, the inclusion of context information allows to correct some mistakes performed when only the appearance was considered (see Figure 3.b). In the third row, the method failed classifying some parts of the *road* as *sky*, while now this confusion is avoided. The information provided by neighboring objects also allows to correctly classify in the last stage of region fusion some small areas of the image which were initially classified as unknown.

4. CONCLUSIONS AND FURTHER WORK

We have presented a probabilistic model for labelling 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 current, to obtain a more accurate result. The labels are in agreement with the image statistics and with the absolute contextual information as well. In the future we will study how to label the objects respect geometric relationships between objects as well as to apply the method in a set of images containing more objects (cars, people, buildings, etc.). Then, we intend to work towards evolving efficient schemes to generate distribution over scene hypothesis using the pixel maps.

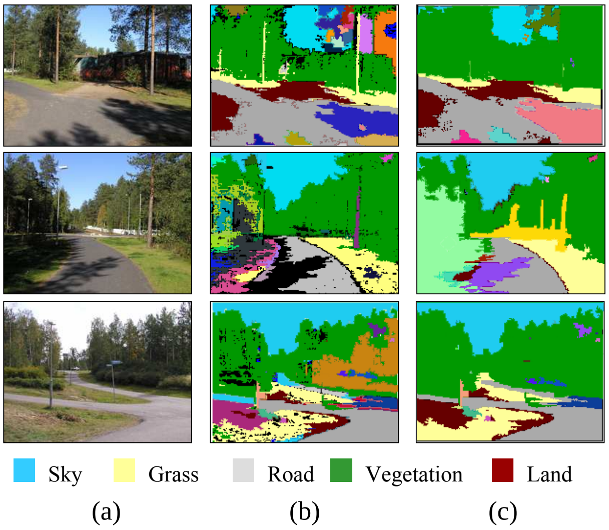

Fig. 4. Experimental results. (a) Original image, (b) appearance-based proposal, and (c) our whole proposal.

5. REFERENCES

[1] 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.

[2] 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.

[3] A. Torralva, “Contextual priming for object detection,” *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[4] A. Singhal, J. Luo, and W. Zhu, “Probabilistic spatial context models for scene content understanding,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, June 2003, vol. 1, pp. 235–241.

[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] J. Martí, J. Freixenet, J. Batlle, and A. Casals, “A new approach to outdoor scene description based on learning and top-down segmentation,” *Image and Vision Computing*, vol. 19, pp. 1041–1055, January 2001.

[7] J. Freixenet, X. Muñoz, J Martí, and X. Lladó, “Color texture segmentation by region-boundary cooperation,” in *European Conference on Computer Vision*, Prague, Czech Republic, May 2004, vol. II, pp. 250–261.

[8] 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.