A Learning Framework for Object Recognition on Image Understanding

Xavier Muñoz, Anna Bosch, Joan Martí, and Joan Espunya

Institute of Informatics and Applications, University of Girona,
Campus de Montilivi s/n 17071 Girona, Spain
{xmunoz,aboschr,joanm,jespunya}@eia.udg.es

Abstract. In this paper an object learning system for image understanding is proposed. The knowledge acquisition system is designed as a supervised learning task, which emphasises the role of the user as teacher of the system and allows to obtain the object description as well as to select the best recognition strategy for each specific object. From several representative examples in training images, an object description is acquired by considering different model representations. Moreover, different recognition strategies are built and applied to obtain initial results. Next, teacher evaluates these results and the system automatically selects the specific strategy which best recognise each object. Experimental results are shown and discussed.

1 Introduction

The aim of image understanding systems is easy to describe. Given an arbitrary photograph, we would like to automatically understand and give meaning to the image by identifying and labeling significant objects in the image. Nevertheless, although we human perform this perception in an immediate and effortless manner, to adequately describe a scene significantly involves the integration of different image processing techniques, pattern recognition algorithms, and artificial intelligence tools, and is a very difficult problem in computer vision [1].

An image understanding system can also be considered as a knowledge-based vision system, because such system requires models that represent prototype objects and scenes. Hence, two important issues must be taken into account: (1) the way in which the model knowledge is organized and stored, and (2) how this knowledge is acquired. However, while knowledge representation has become a permanent focus of interest and a large number of proposals can be found in the literature (see [2]), knowledge acquisition tools are still in their infancy [2].

Early systems were generally not oriented to facilitate the entry of knowledge or carry out some form of automated learning. In contrast, most of the existing systems had to incorporate this new model knowledge by hand; and all the more so for code-encapsulated data. Examples are the Schema system [3] and the region analyzer of Ohta et al. [4], which are successful systems and works of reference. Nevertheless, nowadays most vision researchers agree that the success of scene description systems lies on their ability to learn from experience and

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 311–318, 2005.

training, and there is in last years a clear trend towards the consideration of learning as one of main issues that need to be tackled in designing a visual system for object recognition [5, 6].

Automated learning must consider the acquisition of object models as a description of the object attributes as well as a selection of the strategy used to find and recognize it in an image. Actually, not all objects are defined in terms of the same attributes, and even these attributes may be used in various ways within the matching or interpretation process. Therefore, the learning system must take a flexible and multifaceted recognition strategy into account. A large number of object recognition strategies have been proposed to achieve a particular goal. However, we think is too much pretentious to think that a single method will be able to correctly model and recognize all objects in the real world visual gallery. There is not a perfect strategy for all objects and very little research in the field of computer vision has gone into the problem of determining the best recognition strategies [7].

In this paper we propose an object learning framework for image understanding mainly oriented to outdoor scene images, which addresses the problem of automatic object recognition strategy selection. Inspired on relevance feedback techniques used on image retrieval systems [8], the knowledge acquisition system is designed as a supervised learning task, which involves the user as teacher and part of the learning process. Therefore, the learning allows to obtain the object description as well as to select the best recognition strategy for each specific object under a friendly and effortless user interface.

The remainder of this paper is structured as follows: Section 2 describes the proposed supervised object knowledge learning, focusing on the object description, recognition strategy design and the strategy selection. Experimental results proving the validity of our proposal appear in Section 3. Finally, conclusions are given in Section 4.

2 Learning Proposal

Models constitute representations of the real world, and thus, modeling implies the choice of a suitable set of parameters in order to build those representations. Obviously, the selection of the parameters will affect how well the models fit reality, and this becomes a central issue in any modeling process. Due to the complexity of outdoor scenes, our approach includes the possibility that every single object can be described by specific features and a specific recognition strategy, facilitating later recognition processes. With the aim to provide some improvements in knowledge engineering tasks, system code and models databases become totally independent, with a fast, simple and easy data acquisition process.

Our object learning approach has been designed as a supervised task, which emphasizes the role of the user as the responsible of teaching the system. First, the teacher selects representative examples of objects in training images. From these examples, an object description is acquired by considering different model

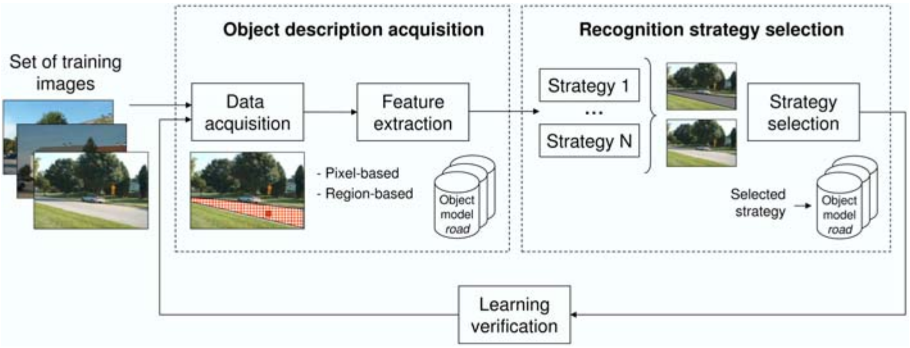

Fig. 1. Scheme of the proposed knowledge acquisition process, working as a supervised learning task.

representations. Next, several strategies to recognize the object are built and applied in order to obtain initial recognition results. These results are evaluated by the teacher, and the system automatically selects the specific strategy which best recognize each object. A global scheme of the proposed knowledge acquisition process is shown in Figure 1.

2.1 Object Description Acquisition

The teacher firstly selects meaningful examples of objects in the training images by clicking on a pixel corresponding to the object of interest. This simple selection allows us to extract the whole object and to compute and register different model representations which provide a complete description of the object. Specifically, the acquired information is composed by:

- **Pixel-based description:** from the selected point, a set of neighboring pixels are extracted and considered as samples of the object pixels. Next, a large number of color and texture features of these pixels are measured. We initially consider the whole set of features as candidates to characterize real objects. In particular, 28 color features related to different color spaces and a set of 8 co-occurrence matrix based texture features are computed for each pixel.
- **Region-based description:** a color texture active region which integrates region and boundary information [9] grows from the selected point in order to segment the region corresponding to the whole object. Color and texture descriptors, as well as shape information based on Fourier descriptors [10], are extracted for each region in order to describe and characterize the object of interest.

2.2 Object Recognition Strategies

The object descriptions can be used in different ways in order to recognize the object. We initially selected, implemented and included into our framework three simple and basic recognition algorithms, which differ on the description model they use as well as the philosophy of the strategy. Recognition strategy A is a top-down approach based on the pixel-level description; recognition strategy B follows a classic bottom-up approach based on a general non-purpose segmentation; while strategy C is a pure hybrid strategy which applies a knowledge-guided search over an initial segmentation. More specifically,

Recognition Strategy A. The top-down strategy consists on the direct search of a specific object by exploiting information concerning the object's characteristics. The implemented method, similarly to the proposal of Dubuisson-Jolly and Gupta [11], models the different objects by a multivariate Gaussian distribution. A pixel-level classification is obtained by using the maximum likelihood classification technique which assigns each pixel to the most probable object.

Recognition Strategy B. A classical bottom-up approach for image understanding was considered for this method. The technique is mainly based on a general purpose segmentation step which tries to part the image into meaningful regions. A color texture segmentation algorithm based on active regions, which integrates region and boundary information [9] was used for this purpose on our implementation. Next, main regions are labeled according to their similarity with stored models.

Recognition Strategy C. Finally, the last implemented method can be considered as a pure hybrid strategy, which starts as the previous approach with a general segmentation. However, a top-down strategy is then performed over these results to specifically find objects of interest. As was noted by Draper et al. [7], not all object classes are defined in terms of the same attributes, and a previous feature selection process allows to select the specific subset of features which best characterizes each single object. Next, selected features are considered to look for the segmented regions on the image which match with the object model.

2.3 Recognition Strategy Selection

Once the process of recognition strategies design is complete, the best specific strategy to recognize each object must be determined. This is the key stage of our proposal; inspired on relevance feedback techniques extensively used on content-based image retrieval systems [8], the role of the user is emphasized and he/she is involved as a vital part of the learning process. With the help of the teacher interaction, the system is able to evaluate the different recognition strategies and to learn which is the best strategy for each object.

Therefore, given a reduced set of training images, the recognition methods are launched together to find all the instances of the given object. Obviously, these strategies can provide different results: because a strategy misses an object apparition, or contrarily it gives a false positive. And in all cases the extraction accuracy must be determined. Hence, these recognition results are visually retrieved to the teacher in order to evaluate their quality. In front of these results, the teacher marks the found instances to indicate if they are well recognized or not. In other words, if the strategy (or strategies) which obtained this recognition was right or wrong. We provide the teacher with three levels of correctness: highly correct, correct and wrong. Although the use of more levels could probably provide more information, we consider it would be lesser friendly for the user to interact with the system. When results have been evaluated by the teacher, this information allows to the system measure the score of each strategy and finally select the strategy which best recognize the object.

The learning process ends with a final verification step. A visual feedback is provided by means of recognizing the object in the set of training images. Obviously, the specific selected strategy will be used for each object. This visual feedback guides the teacher, giving him the option to interfere in the learning process by introducing new training images.

3 Experimental Results

We applied our method to a color image data set constructed using 100 images from the image database of the University of Oulu [12] and also a set of images taken by ourself. These images consists on natural outdoor scenes and mainly contain typical objects in rural and suburban area. We segmented and labeled them manually into 4 classes: *sky*, *grass*, *trees*, *road*, while the remaining areas are considered as *unknown* objects. The training set includes 20 selected images and the remaining 80 were used for testing. We evaluate both, the method selection from the user interaction (section 3.1), and the final goodness of the recognition using the selected strategy (section 3.2). Furthermore, the system is available on an on-line web-based application at *http://ryu.udg.es*.

3.1 Learning Results

The selection of a specific object recognition method from the user feedback and how the system is able to capture the user's criterion, is evaluated by measuring its match with the quantitative results obtained for the different techniques over the training images. Ideally, the selected method must be that which achieves the best results. Table 1 summarizes the scores assigned to each method from the teacher judgements. For example, the user qualifies the top-down strategy for recognizing the road with a quality percentage of 87.50%, which means that the user mostly agrees with the results obtained by this strategy in the recognition of the *road* object. The method which obtains the best score is specifically selected for each object. Summarizing, the top-down strategy A was selected for the recognition of *road* and *grass*, while the hybrid strategy C was considered for the *sky* and *trees*.

On the other hand, Table 2 shows the quantitative evaluation of the strategies over the same training images by measuring the percentage of well classified and over-classified pixels. As was desirable, the strategies selected by the user to classify each object achieve the best results, which means the system is able to capture the user feedback and selects the best method over the set of training images. However, the selection for the *sky* object must be explained. In this case, the system selected the strategy C while the strategy A obtained a 100% of well classified pixels. Nevertheless, this initially wrong decision can be justified by the high percentage of wrongly over-classified pixels which obtains the A top-down strategy. Figure 2 shows the object classification obtained by both techniques over some images of the training set. As is stated in the first column, the top-down strategy A correctly extracts the sky, but confuses some road pixels with this object, while a better recognition is achieved by the hybrid strategy C, which reaffirms the selection performed by our system from the user feedback.

Table 1. Scores obtained to recognize each object taken into account the user criterion. (TD = Top-down; BU = Bottom-up; H = Hybrid.)

Percentage acquired from the user				
Strategy	sky	road	grass	trees
strategy A (TD)	37.50%	87.50%	50.00%	−50.00%
strategy B (BU)	10.00%	20.35%	24.5%	10.18%
strategy C (H)	100.00%	71.43%	35.00%	44.12%

Table 2. Percentages of well classified and over-classified pixels over the training images. (TD = Top-down; BU = Bottom-up; H = Hybrid.)

Percentage acquired from the training classified images								
Strategy	sky		road		grass		trees	
	ok	over	ok	over	ok	over	ok	over
strategy A (TD)	100.00%	10.01%	89.24%	0.25%	87.55%	1.39%	73.36%	9.31%
strategy B (BU)	60.00%	15.00%	60.49%	3.00%	61.04%	7.00%	63.87%	5.00%
strategy C (H)	90.15%	4.38%	75.84%	0.43%	83.13%	3.65%	88.31%	3.13%

3.2 Classification Results

Table 3 summarizes the object recognition results obtained by both selected strategies (top-down strategy A and hybrid strategy C) over the test images set. Moreover, the last row show the percentages obtained by each selected specific object method. The last columns shows the percentages taken into account all objects. From these quantitative results, the significant improvement that is achieved by the use of a specific method for each single object and, the combination of strategies on the whole system in front of a single technique, is stated. Specifically, using the set of selected strategies the system obtains a 85.30% of well-classified pixels, which is clearly superior to the scores obtained by both individual techniques. Moreover, the lowest percentage of wrongly over-classified pixels, 1.88%, is obtained. The results can be qualitatively evaluated in Figure 3,

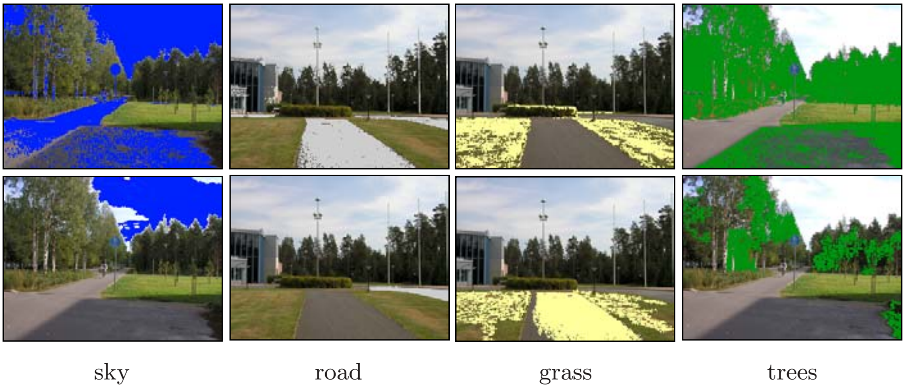

Fig. 2. Some labeling results over the training images set. First row shows results by the top-down strategy A; while second row shows results by the hybrid strategy C.

Table 3. Object classification and over-classification rates over the test images. The last row shows the percentages achieved by the object specific strategy selected from the user feedback.

Percentage acquired from the testing classified images										
Strategy	sky		road		grass		trees		average	
	ok	over	ok	over	ok	over	ok	over	ok	over
strategy A	64.49%	4.94%	86.46%	0.30%	83.20%	1.38%	67.79%	10.69%	75.48%	4.30%
strategy C	83.92%	0.37%	67.24%	0.33%	46.06%	2.27%	87.62%	5.51%	71.21%	2.80%
selected	83.92%	3.15%	86.46%	0.30%	83.20%	1.38%	87.62%	5.51%	85.30%	1.88%

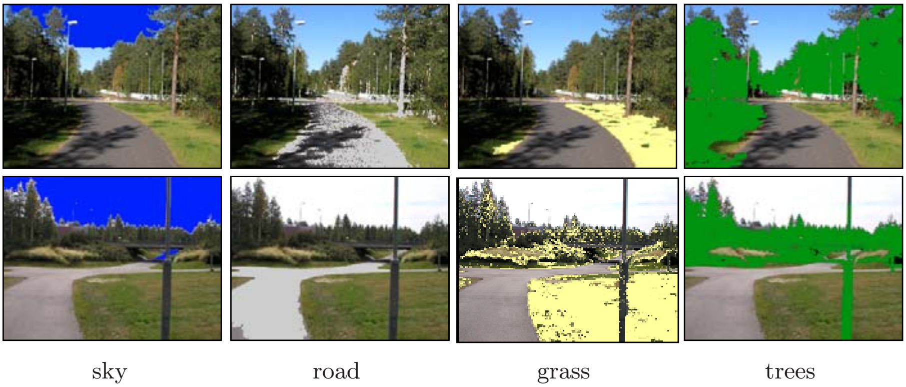

Fig. 3. Some labeling results over the test images set using by each object the strategy selected by the proposed learning framework.

which shows the object recognition achieved by our system using specifically selected object recognition techniques, and denotes the correctness of the object learning and extraction.

4 Conclusions

An object learning framework for image understanding has been described. The process has been designed as a supervised learning task, which emphasizes the role of the user as system teacher. From some examples provided by the teacher, the system extracts the information required to describe the object. Moreover, the learning allows to select the best recognition strategy for each specific object under a friendly and effortless user interface. Experimental results has stated the convenience of using a set of object specific recognition methods.

Extensions of this work are oriented to the improvement of the strategy selection to make possible the combination of several techniques as the best method to recognize an object. Furthermore, new recognition strategies will be included into the system.

References

1. Yun-tao, Q.: Image interpretation with fuzzy-graph based genetic algorithm. In: IEEE International Conference on Image Processing, Kobe, Japan (1999) 545–549
2. Crevier, D., Lepage, R.: Knowledge-based image understanding systems: A survey. Computer Vision and Image Understanding **67** (1997) 161–185
3. Draper, B., Collins, R., Brolio, J., Hanson, A., Riseman, E.: The schema system. International Journal of Computer Vision **2** (1989) 209–250
4. Ohta, Y.: Knowledge-based Interpretation of Outdoor Natural Color Scenes. Pitman Publishing, Boston, Massachussets (1985)
5. Drummond, T.: Learning task-specific object recognition and scene understanding. Computer Vision and Image Understanding **80** (2000) 315–348
6. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 2. (2003) 264–271
7. Draper, B., Hanson, A., Riseman, E.: Knowledge-directed vision: Control, learning, and integration. Proceedings of the IEEE **84** (1996) 1625–1637
8. Rui, Y., Huang, T.S. Ortega, M., Mehrotra, S.: Relevance feddback: A power tool for interactive content-based image retrieval. IEEE Trans on Circuits and Systems for Video Technology **8** (1998) 644–655
9. Muñoz, X., Freixenet, J., Cufí, X., Martí, J.: Active regions for colour texture segmentation integrating region and boundary information. In: IEEE International Conference on Image Processing, Barcelona, Spain (2003)
10. Zhang, D., Lu, G.: Generic fourier descriptor for shape-based image retrieval. In: IEEE International Conference on Multimedia and Expo. (2002)
11. Dubuisson-Jolly, M.P., Gupta, A.: Color and texture fusion: Application to aerial image segmentation and gis updating. Image and Vision Computing **18** (2000) 823–832
12. Ojala, T., Mäenpää, T., Pietikäinen, M., Viertola, J., Kyllönen, J., Huovinen, S.: Outex - new framework for empirical evaluation of texture analysis algorithms. In: IAPR International Conference on Pattern Recognition. Volume 1. (2002) 701–706