# Combining neural networks and clustering techniques for object recognition in indoor video sequences

Francesc Serratosa[1], Nicolás Amézquita Gómez[1] and René Alquézar[2]

[1]Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Campus Sescelades, Av. dels Països Catalans 26, 43007, Tarragona, Spain
[2]Dept. Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, 08034 Barcelona, Spain
{francesc.serratosa, amezquita}@urv.cat, alquezar@lsi.upc.edu

**Abstract.** This paper presents the results obtained in a real experiment for object recognition in a sequence of images captured by a mobile robot in an indoor environment. Objects are simply represented as an unstructured set of spots (image regions) for each frame, which are obtained from the result of an image segmentation algorithm applied on the whole sequence. In a previous work, neural networks were used to classify the spots independently as belonging to one of the objects of interest or the background from different spot features (color, size and invariant moments). In this work, clustering techniques are applied afterwards taking into account both the neural net outputs (class probabilities) and geometrical data (spot mass centers). In this way, context information is exploited to improve the classification performance. The experimental results of this combined approach are quite promising and better than the ones obtained using only the neural nets.

**Keywords:** Clustering, Spot, Class probabilities.

## 1. Introduction

One of the most general and challenging problems a mobile robot has to confront is to identify and locate objects that are common in its environment. To this end, a teacher may show what the object is from images taken in different views and a robot may apply some learning abilities to obtain a certain model of the object and an associated recognition procedure.

A very important issue is to determine the type of object model to learn. In our point of view, a useful model should be relatively simple and easy to acquire from the result of image processing steps. For instance, the result of a color image segmentation process, consisting of a set of regions (spots, from now on) characterized by different features (related to color, size and shape), may be a good starting point to learn the model. Although structured models like adjacency attributed graphs or random graphs can be synthesized for each object from several segmented images [1], we have decided to investigate first a much simpler approach in which the object is just

represented as an unstructured set of spots. One of the main drawbacks of the structural methods is that the segmented images from one frame to the other can be quite different, and so, it is difficult to match the actual spots (usually represented by nodes of the graphs) with the previous ones. The main aim of our approach is to accept these differences between segmented images and use a more coarse approach in which the basic element is not the spot or region of the segmented image but its pixels.
In a previous recent work [2], feed-forward neural networks were used to classify the spots independently as belonging to one of a finite set of objects or the background (defined as everything else). In this work, clustering techniques are applied afterwards taking into account both the neural net outputs (class probabilities) and geometrical data (spot mass centers). In this way, context information can be exploited to improve the classification performance.

The classification of segmented image regions for object recognition has been addressed in several works. Neural networks are used in [3] not only to classify known objects but to detect new image objects as well in video sequences. In [4], objects of interest are first localized, then features are extracted from the regions of interest and finally a neural network is applied to classify the objects. Support vector machines are used in [5] to classify a segmented image region in two categories, either a single object region or a mixture of background and foreground (multiple object region), in order to derive a top-down segmentation method.

## 2. Image acquisition, pre-processing, segmentation and feature extraction

The input of our system is a digital video sequence. The images in the sequence were preprocessed by applying a median filter on the RGB planes and segmented by the Felzenszwalb – Huttenlocher algorithm [6]. The output of the segmentation and feature extraction process for each image consists of a list of spots (that represent regions) with their features.

Two types of information were extracted from the spots: color and geometry. With regards to color, average and variance values for each one of the three RGB bands were calculated for each spot on the basis of the corresponding intensity values of the spot pixels in the original image. This is, the result of the segmentation algorithm served to identify the pixels of every spot, but their color features were computed from the original RGB image.

Regarding the geometrical information, we were mainly interested in shape descriptors that were invariant to translation and scale, and to this end, we decided to use the seven invariant geometric moments defined by Hu [7] (whose equations are also reproduced in [2]). In addition and since the range of variation of the objects' size was rather limited in the video sequence, we also calculated and used the size of each spot, i.e. its area measured in number of pixels. Hence, 14 features (6 for color

and 8 for geometry) were computed for each spot. Moreover, the mass center was also calculated.

## 3. Spot classification methodology

A neural net is first trained to classify each spot within certain regions of the images in the sequence. The inputs of the neural net are the features of each spot and the target is the class that we impose to each spot. To impose the class of each spot, we manually marked on the images a rectangular box for each object of interest. Thus, the spots whose mass centers are inside each box are forced to belong to the class that the box represents. Figure 1 shows one of the images and its segmentation together with the boxes used to impose the classification.

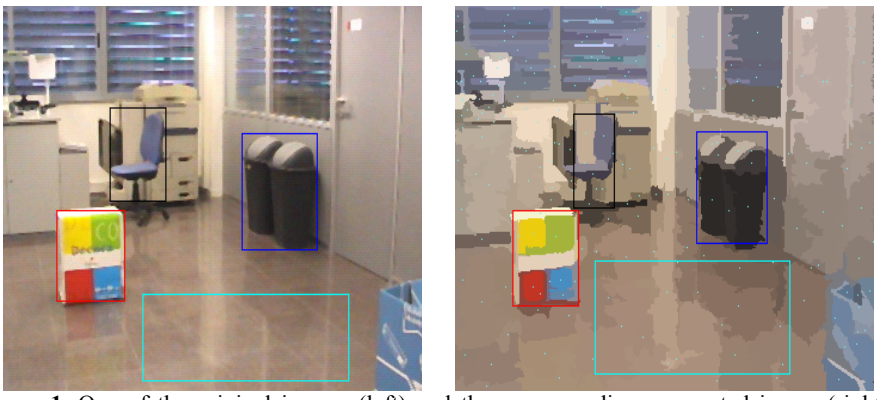

**Figure 1.** One of the original images (left) and the corresponding segmented image (right), with four boxes marked on them. Spot mass centers are also displayed in the right image.

In the recognition stage, each image from the video sequence is segmented, then for each spot of the segmented image, a first classification was made using the neural net. A second step for spot classification involves the detection and reclassification of possibly misclassified spots based on the context information provided by the mass centers of the spots classified as the same class (or object) in the same frame.
For each one of the classes (or objects) o and for each frame f, a weighted mass center wmc(o,f) was computed as

$$wmc(o,f) = \frac{\sum_{s=1}^{ns(o,f)} p(o/s)a(s)mc(s)}{\sum_{s=1}^{ns(o,f)} p(o/s)a(s)} \quad \textbf{(1)}$$

where ns(o,f) is the number of spots classified as object o in frame f, p(o|s) is the a-posteriori class probability of object o for spot s given by the net, and a(s) and mc(s) are respectively the area and mass center of s.
Then, for every spot in the segmented image classified by the net as an object, the distance between its mass center and the weighted mass center of the assigned object was computed. If this distance exceeded a given threshold, the spot was marked as possibly misclassified and it was optionally reclassified to the object with the nearest mass center. Note that this step is a kind of spot clustering process that is inspired in both the dynamic and k-means clustering algorithms, but starting from the clusters (class assignments) given by the neural network.

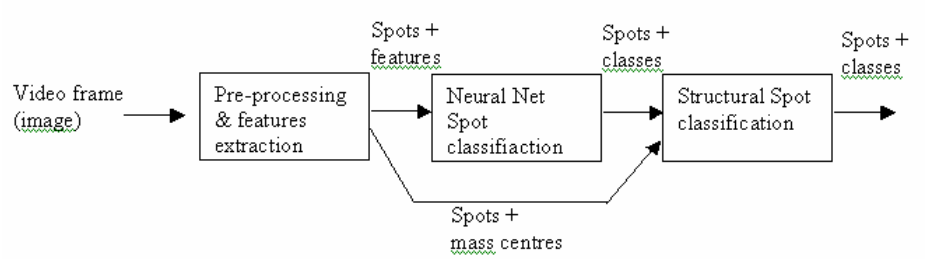


**Figure 2.** Classification process based on three main steps. First, the extraction of the spots and features. Second, spot classification based on a Neural Net and third, reclassification of the spots based on its position respect the other spots within the same class.

Figure 3 displays an example of the beneficial effects of performing the reclassification based on structural information. In the left hand image, there are two spots that were misclassified by the net, one in the chair was classified as wastebasket and one in the wastebasket was classified as chair. These spots could be correctly reclassified after this step, as shown in dark green and dark blue in the right hand image.

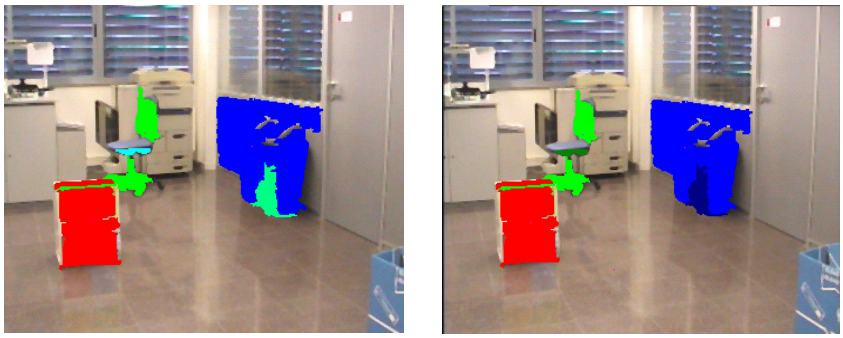

**Figure 3.** Spots classified as belonging to the three objects by the net (left) and the result of the reclassification after the clustering (right).

## 4. Experimental results.

A digital video sequence of 88 images was captured by an RGB camera installed on the MARCO mobile robot at the Institute of Robotics and Industrial Informatics (IRI, UPC-CSIC) in Barcelona. The sequence shows an indoor scene with some slight perspective and scale changes caused by the movement of the robot while navigating through a room. The objects of interest in the scene were a box, a chair and a pair of identical wastebaskets put together side by side (see Figure 1), and the objective was to discriminate them from the rest of the scene (background) and locate them in the images.

For the training stage, four rectangular boxes were manually marked on the images with a graphics device to encompass the three objects of interest and a large region on the floor (figure 1). In order to assign a class label to each spot, to be used as target for the spot pattern in the neural network training and test processes, a simple decision was made: each one of the four rectangular boxes constituted a class and all the spots that had its mass center inside the window and a size large than 100 pixels were assigned the same class label.

We used a neural net with a feed-forward 2-layer perceptron architecture using standard backpropagation as training algorithm. From previous experiments reported in [2], we set the number of hidden units to 180, although it was shown in [2] that the results were not very sensitive to this choice. Hyperbolic tangent and sine functions were used as activation functions in the hidden layer and the output layer, respectively. For backpropagation, we set a learning rate of 0.003, a momentum parameter of zero and a maximum number of 500 training epochs for each run.

A dataset containing 3,411 labeled patterns (spots) was available after the segmentation of all the 88 images. For each subset of features, a double cross-validation procedure was carried out that generated 90 different partitions of this dataset, each including 80% of the patterns for the training set, 10% for the validation set and 10% for the test set. The validation sets were used for early stopping the training phase. Actually, the network chosen at the end of the training phase was the one that yielded the best classification result on the validation set among the networks obtained after each training epoch.

The results of the double cross-validation procedure obtained for different subsets of features are displayed in Table 1, ordered decreasingly by test classification performance. For each one of the three sets (training, validation and test set), the classification performance is measured as the average percentage of correctly classified patterns in the 90 cross-validation partitions, evaluated in the networks selected after training (the ones that maximize the performance on the validation set). It can be noted that similar results are obtained if the average RGB color features are taken into account, but the performance falls down dramatically when they are not used. The best result was 92.93% test classification performance for a subset comprising color features (both RGB averages and variances) and spot size (and without the shape invariant moments).

These results are in agreement with those reported in [2] with regards to the relative usefulness of the different spot features (i.e., invariant moments are shown to be practically useless and RGB color averages are shown to provide almost all the relevant information), but the absolute classification rates are notably better here, due to a more accurate definition of the rectangular boxes that eliminated from the dataset most of the spots that were incorrectly labeled in [2].

**Classification performance (with feature subsets)**

| Feature Subsets | Training | Validation | Test |
|---|---|---|---|
| spot size, average and variance RGB | 94.20 | 93.38 | 92.93 |
| spot size and average RGB | 93.29 | 93.26 | 92.75 |
| all 14 features | 94.58 | 93.19 | 92.72 |
| spot size, average RGB and three first moments | 93.47 | 92.92 | 92.22 |
| Average RGB and three first invariant moments | 92.11 | 92.37 | 91.93 |
| Average RGB and the seven invariant moments | 92.04 | 92.35 | 91.60 |
| spot size and variance RGB | 62.12 | 63.54 | 63.06 |
| spot size, variance RGB and three first moments | 62.46 | 63.52 | 62.79 |
| seven invariant moments and variance RGB | 55.98 | 57.48 | 57.33 |
| seven invariant moments | 32.29 | 32.49 | 32.38 |

**Table 1.** It presents the classification results for several groups of selected variables to assess the relative importance of the different types of features (size, color averages, color variances and shape invariant moments).

Using spot size and RGB averages and variances as features, the network (and associated dataset partition) that gave the best result in the training set (97.25%) was selected for computing the weighted mass centers and to assess the effect of the clustering process on the spot classification performance. Table 2 compares the results obtained without clustering with those obtained after clustering and reclassification to the nearest object. A 78.8% of the spots misclassified by the network were correctly reclassified by the clustering and only a 0.1% of the correctly classified spots were incorrectly reclassified.

**Classification performance (with the best feature subsets)**

| Classifier | Training | Validation | Test |
|---|---|---|---|
| Only the neural network | 97.25 | 95.01 | 96.18 |
| Combining the neural net and the clustering | 99.34 | 98.53 | 99.71 |

**Table 2.** Spot classification results before and after clustering using the net that maximized the result in the training set.

Figure 3 displays an example of the beneficial effects of performing the structural reclassification. In the left hand image, there are two spots that were misclassified by

the net, one in the chair was classified as wastebasket and one in the wastebasket was classified as chair. These spots could be correctly reclassified after the structural reclassification, as shown in the right hand image.

## 5. Conclusions and future work

A simple approach to object recognition in video sequences has been successfully tested combining neural networks and clustering techniques to classify image segmentation regions (spots) as belonging to one of the objects of interest or to the background. Objects are implicitly represented as an unstructured set of spots; no adjacency graph or description of the structure of the object is used. The method is robust to changes between successive frames in the number and shape of the spots associated with each object, as given by the image segmentation algorithm.

In this work, spatial context information has been obtained through the distances between the mass centers of the spots, which allow the formation of semi-supervised clusters, since both the classification labels and probabilities given by the neural net are taken into account as well for the clustering. Other ways of aggregating spatial context can be studied in future work, e.g. relaxation labeling may be used for updating the class probabilities of neighboring spots.

The obtained classification results are quite good, but it must be noted that only the spots in some regions of interest were processed and just three objects (plus background) were considered as classes. A more realistic experiment would involve the spots of whole images in the test phase and eventually more objects to recognize. Moreover, the dynamic nature of the visual data should be exploited by somehow integrating the tasks of object detection, recognition and tracking in consecutive video frames.

In the long-term, our purpose is to design a robust dynamic approach to object recognition and tracking in video sequences based on unstructured sets of spots, which can deal with the variations in the object views resulting from the movement of a mobile robot in an indoor environment.

## 6. References


1. Sanfeliu A., Serratosa F., Alquézar R., “Second-order random graphs for modeling sets of attributed graphs and their application to object learning and recognition”, Int. Journal of Pattern Recognition and Artificial Intelligence, Vol. 18 (3), 375-396, (2004).
2. Amezquita Gómez N. and Alquézar R. “Object recognition in indoor video sequences by classifying image segmentation regions using neural networks”, Proc. 10th Iberoamerican Congress on Pattern Recognition, CIARP 2005, Havana, Cuba, M. Lazo and A. Sanfeliu (eds.), Springer-Verlag, LNCS 3773, Berlin, pp.93-102, (2005).
3. Singh S., Markou M., Haddon J., “Detection of new image objects in video sequences using neural networks”, Proc. SPIE Vol. 3962, pp.204-213, Applications of Artificial Neural

Networks in Image Processing V, Nasser M. Nasrabadi; Aggelos K. Katsaggelos; Eds., (2000).
4. Fay R., Kaufmann U., Schwenker F., Palm G., “Learning object recognition in a neurobotic system”. In: H-M. Groß, K. Debes, H-J. Böhme (Eds.) 3rd Workshop on SelfOrganization of AdaptiVE Behavior (SOAVE 2004). Fortschritt -Berichte VDI, Reihe 10 Informatik / Kommunikation, Nr. 743, pp. 198-209, VDI Verlag, Düsseldorf, (2004).
5. Wang W., Zhang A. and Song Y., “Identification of objects from image regions”, IEEE Int. Conf. on Multimedia and Expo (ICME 2003), Baltimore, July 6-9,(2003).
6. Felzenszwalb P. and Huttenlocher D., “Efficiently computing a good segmentation”. In IEEE Conference on Computer Vision and Pattern Recognition, 98-104, (1998).
7. Hu M-K., “Visual pattern recognition by moment invariants”, IRE Trans. on Information Theory, Vol. 8 (2), pp. 179-187, (1962).
8. Bishop C.M., Neural Networks for Pattern Recognition, Oxford University Press,(1995).