Complex Object Recognition Using a Biologically Plausible Neural Model

Dipl. Eng. RAUL MUREŞAN
Nivis Research
Gh. Bilascu, Nr. 85, Cluj-Napoca, Cod 3400
ROMANIA

Abstract: - The complex object recognition tasks are still one big problem in neurocomputing today. This paper presents a method of detecting and recognizing complex objects, in cluttered environment, in a purely feed-forward way, being able to account for ultra-rapid visual categorization. We used a retinotopic architecture of simple spiking neurons with different types of receptive fields, organized in a hierarchical fashion similar to the mammal visual path. Fast shunting inhibition had been implemented using a rank-order coding similar to that described by S. Thorpe. The main advantage of the neural model proposed is that it accepts a very small number of training examples (4-7) being able to generalize very well. The model has been used to detect faces and automobiles in complex intensity images.

Key-words: - Scale independence; Rank order coding; Feed-forward; Receptive field; Neurocomputing.

1. Introduction

Numerous studies on the mammal visual cortex had shown a general structure and architecture, which can account for the ability to recognize objects. The biological model is a hierarchical one, each layer having a well-determined structure and function.

Most of the models created before 1990 were based upon the assumption that neurons transmit information using pulse rate coding. These models may be consistent, but they have one big problem: they are too slow to be realistic and biologically plausible in rapid visual categorization.

Recent studies however, had shown that ultra-rapid visual categorization is possible, in a time magnitude under 150 ms in human visual neocortex. This is exactly the timing necessary for the information to reach the infero-temporal (IT) cortex neurons responsible for object recognition [6].

Under such circumstances, a natural question arises: how can the brain recognize objects (mainly unfamiliar objects, presented only few times to subjects) in a scale and position independent manner since there is no time for pulse synchronization to occur?

As well, a very important observation is that usually realistic neural systems such as the neocortex do not require a long training phase before being able to generalize in a powerful manner. The second natural question is how can the brain achieve such generalization performance under such a short period of training (visual analysis)?

In the next sections we describe a general model based on the biological ventral visual path. This general model had been used to recognize complex objects from intensity images, under the constraint of very small number of training examples (4-7).

2. Methods

For testing reasons and modeling, we implemented a neural simulator based on the retinotopic organization of the visual cortex. The simulator, named "RetinotopicNET" can successfully trace networks with millions of neurons and a magnitude of 10^{10} synapses in a matter of seconds. This high performance is due to the event-based type of simulation.

Neurons were simple integrate-and-fire cells with fast shunting inhibition implemented as exponential modulated synapses [7]. No leakage has been included in the model since the amount of current leak, in the short period the neuron's state is pooled, can be neglected (no rate based coding is present).

Each time an afferent neuron spiked, the sensitivity of the synapses had been decreased by a modulation factor as follows:

$$\text{Sensitivity} \leftarrow \text{Sensitivity} * \text{Modulation}$$

where:

- Sensitivity represents the synaptic sensitivity over all synapses
- Modulation is a number in the range [0..1]

The architecture of the model contains 6 levels of processing, following the retinal, V1, V4 pathway up to the infero-temporal cortex.

2.1. Architecture

The six layers of processing correspond to an ascending feed forward processing with lateral interactions at some of the levels (Figure 1). The only information used at this time is contour information but blob-type cells could also be included to account for color or intensity patches, as well.

Level 1 : Retinal processing

At the first layer of processing the retinal ganglion cells process the incoming image intensities (only 8 bit grayscale images were used). The ON-OFF effect has been achieved by using a classical difference-of-gaussians (DOG), center-ON-surround-OFF and viceversa filter with a ratio of 1 to 3. Then, the image intensity for the two maps has been converted into spike latency and spikes were fed into the "RetinotopicNET" simulator.

Level 2 : V1 Area

The second layer of processing corresponds to the V1 primary cortex area where a model of oriented contrasts is created. Oriented Gabor-like receptive fields select orientations. These are the corresponding simple cells, which detect different orientation contrasts.

One key feature is the lateral connection within each orientation map. We have used a butterfly-like lateral connection, which has the property of improving contours. This is a form of primitive contour-integration, but due to the lack of iterative loops only a feed forward contour completion is achieved. Important work on this matter had been conducted by Zhaoping Li [3]. Further improvement on the system may be achieved by implementing a stronger contour integration mechanism.

The Gabor patches were all at the same scale and had a spatial frequency of 0.5 pixels. They covered the range of 0 to 180 degrees spanning over a total of 8 orientations.

Level 3 : Multiscale downsample

The third layer of processing is responsible for bringing every detail to the same level of spatial importance. We used here a simple neural scaling model. Scaling is achieved by different sizes of the receptive fields. The receptive fields are not overlapping and the neurons in this layer act as 'OR' functions. This type of behavior is due to the synaptic weight, which is sufficient to determine that a single afferent spike can drive the neuron into the critical area.

Five levels of scales had been used, a larger number offering a better scale independence (1:1 to 1:1.68).

Different strategies are used in learning and matching operations. Learning is supervised and an appropriate scale map is selected by specifying one of the object's dimensions. The rest of the scales are inactive during learning. In the process of recognition, the maps are sequentially activated until all the scales had been tried or the target recognized (final map fired).

The neural downsampling is achieved by using window-like receptive fields which could be associated with center-ON surround-OFF receptive fields in area V4, taking into account the fact that the surround-OFF is a very silent small inhibition which could be used for stability and normalization purposes.

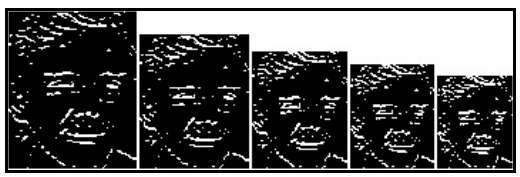

Fig. 2. Multiscale neural downsampling for 5 levels of scale

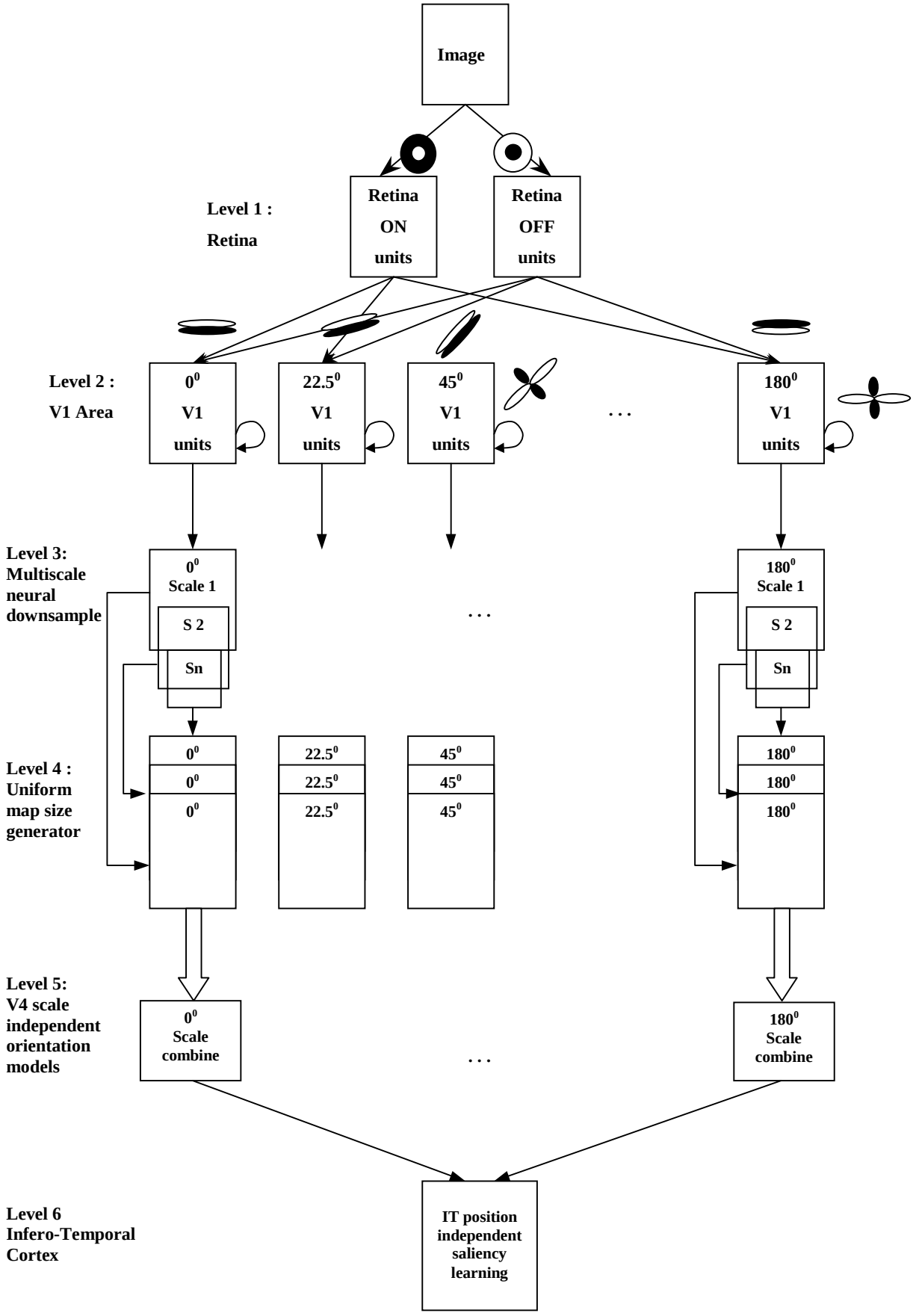

Fig. 1. Model architecture

Level 4 : Uniform map size

At the next layer, at each orientation, the downsampled-oriented maps are brought to the same size. The size should be the same with the one for the next layers because, after selecting the scale (for learning or matching) the model of scaled orientations should only be committed to stimulus property. In this way, the next level can treat all the scaled versions uniformly. Moreover, the appropriate positioning in the original image is done at this level. This is because the original object should be correctly positioned, even after being scaled to a standard size.

Level 5 : Independent orientations

The fifth layer is used for adaptive selection of scales. The levels 3 to 5 achieve a uniform scale, at the last level. For this particular reason, in learning, one of the 5 scales is gated from the fourth level and drives the corresponding neural map at the fifth level. The mechanism described allows different scaled original stimulus to be learned at the same final scale (scale independent learning).

During recognition, the five scales are gated in a sequential way, until one of the scales best matches the standard scale learned. By analyzing the disproportion between standard scale and the map gated from the previous stage, one can even determine the scale of the object in the original image.

Level 6 : Infero-Temporal Cortex

The infero-temporal cortex is responsible for object recognition. In the architecture presented, learning is performed by increasing the synapse strength with the current sensitivity value as resulted from successive modulator effects generated by the firings in the level 5 map. This mechanism is similar to the one used by Arnaud Delorme [1]. Every neuron in the final IT map has a retinotopic type of receptive field, covering most of the level 5 map. The synaptic strengths are shared among all neurons, yielding good position independence.

2.2. Learning

Supervised reinforcement learning is done at the last level of the model (infero-temporal cortex). Two types of information are used for training purpose: position and scale. The user should select three main points on the original object (the upper corners and the bottom center point). Using these 3 points, the center of the triangle is computed and considered the center of the object. Furthermore the distance between the upper two points is compared to a standard, fixed size. The comparison process yields the relative scaling necessary for achieving standard scale at the fifth level of the model. Based on this information, the selection of the appropriate fourth level map is done.

Position is used to determine the neuron that should be trained in the infero-temporal map. After selecting such a neuron, reinforcement learning is done using the instantaneous sensitivity level of the neuron.

2.3. Recognition

The process of recognition has to determine three types of information: the object's presence or absence, then its position and finally its scale.

Recognition starts with propagating the spikes in a feed-forward fashion and monitoring the infero-temporal spikes. One spike in the infero-temporal map is associated with target recognition (target hit). The position of the neuron that spiked corresponds to the position of the object's center (as defined by the three points selected on learning).

At the fifth level, for each trial, one scaled map is selected from below and gated to drive the fifth level map. The selection is done in a sequential way. On a spike event in the IT map, scale can be determined using the gated map's size.

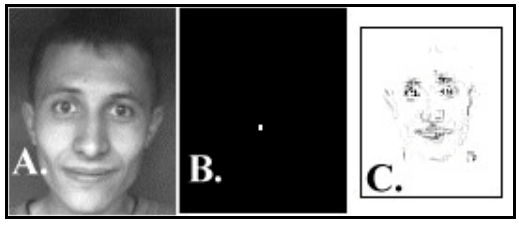

Fig. 3. Face recognition. A. Original image. B. Infero-temporal spike. C. Selectivity of intero-temporal neurons.

3. Results

Using the "RetinotopicNET" simulator we calibrated the system for face detection (and recognition) and vehicle detection. The test database had been generated using a QuickCam web camera and consisted of the faces of 43

different persons with different face expression. The ORL face database had also been included for testing.

Image sizes were fixed at 92 x 112 grayscale bitmaps and the faces were scaled in a range 1:1 to 1:1.58 the original scale.

At the infero-temporal level, a strong shunting inhibition had been used to provide enhanced selectivity on learning (for the face recognition case).

The first objective was to determine the ability of the model to generalize after only a small number of training examples. It proved that important generalization capability emerged after only 3 to 5 training examples for the face identification case and 5 to 7 for the vehicle detection problem.

The capability of generalization has been tested using only 5 training examples and then testing against 18 novel targets (face identification case). The medium activation of IT neurons has been determined (the model uses a resting potential of -65 mV and a firing threshold of -45 mV) (see Fig. 4).

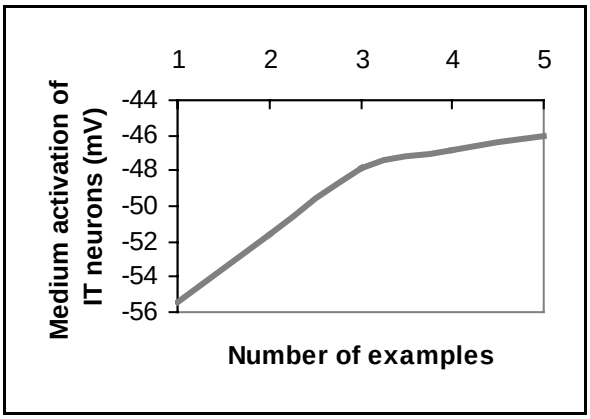

Fig. 4. Medium activation of IT neurons over 18 novel targets for different number of training sessions.

The next step was to determine the ability of the model to recognize faces. The training set of examples had been mostly chosen in a random fashion, using only 4 examples (3 of them randomly chosen and 1 as being the next most distant image - in terms of IT activity) (Fig. 5).

The non-target database consisted of 40 test images of 27 persons.

For the face identification task, the model made only 1 mistake (false prediction) and 2 recognition misses (one due to interscale size and one because of non-uniform environment).

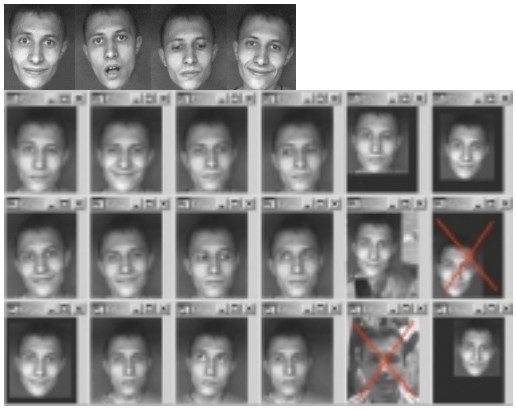

Fig. 5. Train examples (upper 4) and novel target set of images.

The required background environment for correct identification hadn't been used in one case, this leading to one additional target miss (the first is due to scale). The interscale miss can be eliminated very easily, by increasing the scale domain.

Images	No.	True pos.	True neg.	False pos.	False neg.
Target	18	16	-	-	2
Non-target	40	-	39	1	-

Table 1. Testing results for the face identification task.

The results were analyzed and the general accuracy, sensibility and specificity of the method (on the database used) had been determined.

G.A.	SE.	SP.
94.8%	88.8%	97.5%

Table 2. General accuracy, sensibility and specificity for the face identification task.

As mentioned, the face identification task requires a uniform background for correct identification. We tested then the ability of recognizing complex objects from cluttered environment, disregarding the background. For this purpose car recognition task has been developed.

For a good generalization, a set of 7 training examples had been presented to the system. The target images were 19 frames from a highway-recorded video.

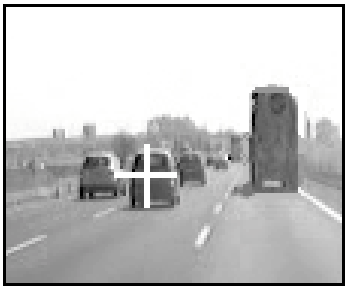

Fig. 6. Car detection at a fixed scale.

The car detection process took no advantage of scale independence, using only one scale (for speed purpose) and only a fixed scale car had been considered as target.

From 19 frames, the system made only one mistake, due to the lack of scale independence. On a Pentium® 4 processor, recognition had been achieved under 400 milliseconds.

The given results are surprisingly good, taking into account the limited scale levels used. Increasing the number of the scale levels (to cover a wider field of scales) can increase accuracy of recognition. At the same time, we expect that the usage of more orientations can increase accuracy because of the better localized detection at the second level of the architecture.

Complex objects can be recognized using this general architecture and no background constraint is necessary for general visual categorization.

The participation to the Conference was supported by Nivis (www.nivis.com).

References:

[1] **Delorme, A., & Thorpe, S.,** Face recognition using one spike per neuron: resistance to image degradation. *Neural Network, in press,* 2001.

[2] **Hubel, D.** & **Wiesel, T.,** Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.,* No. 28, 1965, pp. 229–289.

[3] **Li, Z.,** A neural model of contour integration in the primary visual cortex. *Neural Comput,* Vol. 4, No. 10, 1998, pp. 903-40.

[4] **von der Malsburg, C.,** The What and Why of Binding: The Modeler's Perspective, *Neuron,* Vol. 24, 1999, pp. 95-104.

[5] **Riesenhuber, M. & Poggio, T.,** Hierarchical models of object recognition in cortex, *Nature Neuroscience,* Vol. 20, No. 11, 1999, pp. 1019-1025.

[6] **Thorpe, S., Fize, D.** & **Marlot, C.,** Speed of processing in the human visual system, *Nature, 381*(6582), 1996, pp. 520-522.

[7] **Thorpe, S.J.** & **Gautrais, J.,** Rank order coding. In J. Bower, *Computational neuroscience: Trends in research,* (pp. 113-118), New-York: Plenum Press, 1998.