Object Detection and Recognition in a Multiresolution Representation

Evangelia Micheli-Tzanakou and Ivan Marsic
Department of Biomedical Engineering, Rutgers University
Piscataway, NJ 08855-0909

Abstract. This paper is concerned with the detection and recognition of visual objects. We are dealing with recognizing objects in natural scenes, not only objects isolated on empty background and preprocessed to various transforms. Two assumptions are vital in the whole process: (a) the object representation is equivalent to some two-dimensional function, and (b) the stability and sensitivity of the representation are preserved and are of equal importance. A truncated multiresolution pyramid is used. The method for object detection is based on neurophysiological findings.

1 Introduction

Detection and recognition of the objects in a visual environment is something that biological visual systems do extraordinarily easily and well. Even the best contemporary artificial neural networks for pattern recognition have serious constraints on the types of objects and scenes they can deal with. Object recognition is a difficult problem because the relationship between the objects in their environment and the image created on the input array is highly variable. The image undergoes various transforms like scaling, translation, rotation, perspective projection, and occlusion.

Most of the research effort is done with the hope that there are some general principles common to all recognition systems. In a recent review paper, Felleman and Van Essen [4] point out that although somatosensory cortex has much fewer areas (13 identified until now) and pathways connecting them (62) than the visual cortex (32 and 305, respectively), the number of hierarchical levels is approximately the same (9 and 10, respectively). This can only mean that the recognition task requires a certain number of steps, which cannot be performed in parallel. The visual system has more parallel pathways only because the visual information has more qualities.

Different theories of shape recognition make different assumptions about the long-term memory representations of objects. Here we assume that the object representation is pictorial, i.e. it is equivalent to some two-dimensional function, which is also supported by recent psychophysical findings [3].

Among the most important considerations about the representation are its *stability* and *sensitivity*. The stability is necessary to avoid any irrelevant variations influence the result of recognition while the sensitivity is needed in order to differentiate between two objects of the same class.

The stability and sensitivity properties of a representation are in contradiction with one another, and therefore they must be implemented through different mechanisms. We use the wavelet transform in order to get another image-like representation in a multiresolution environment.

In order to deal with arbitrary scenes, we have to be able to select the potential objects from the image before matching them to memorized descriptions. Figure/ground segregation also called image segmentation was shown to be a very tough problem in computer vision. We have a rather less ambitious goal, which is just selection of candidates for the matching process. The

CH 3065-0/91/0000-2564 $1.00 © IEEE

object detection mechanism we use is based on neurophysiological findings. The purpose of this mechanism is to reveal the globally salient structures which are candidates for matching with memorized or learned object descriptions.

2 Methods

2.1 Multiresolution Representation

The neural networks for classification usually do not address the question of sensitivity separately. Besides reducing the stability and memory capacity, addressing sensitivity at the level of whole objects has no meaning because sensitivity deals with the local properties of the objects. A natural way to achieve both stability and sensitivity in a representation is through a multiresolution architecture, for example by the wavelet transform [8, 1]. Different scales of detail in the image are best described by appropriate spatial frequencies in the Fourier transform of the image. Different frequency bands contain different types of information in the image.

We have chosen the wavelet transform to produce the multiresolution representation of the input image. The wavelet transform of a 2-D function $f(x,y)$ at the scale s and a point (u,v) is defined by

$$Wf(s,(u,v)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)s\Psi(s(x-u),s(y-v))dxdy. \quad (1)$$

The function $\Psi(x,y)$ can be interpreted as the impulse response of a band-pass filter having no preferential spatial orientation.

Having this representation, we deduced the properties of recognition processes, operating on this representation [9]. The multiresolution architecture is shown in Fig.1.

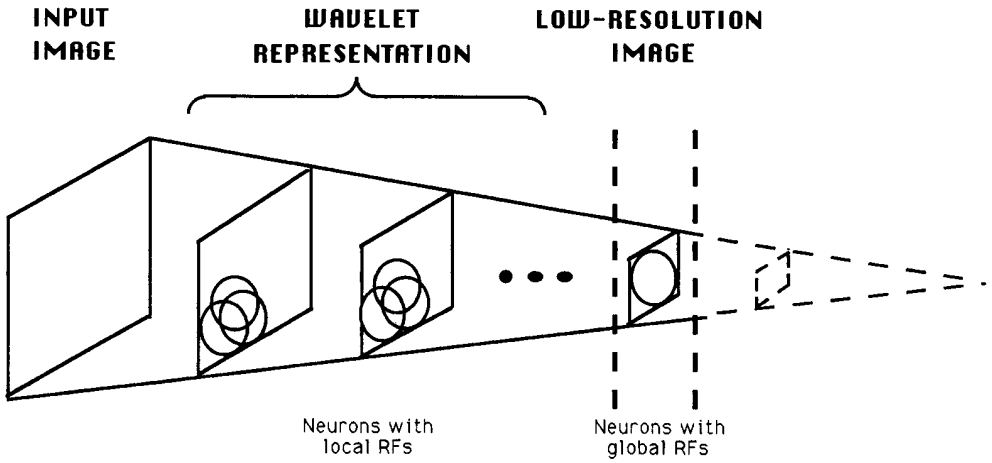

Figure 1: The truncated multiresolution pyramid architecture, which allows for stability and sensitivity of the representation. Of the neurons performing template matching, only one layer (lowest resolution) has global RFs. The neurons at higher resolution have local RFs.

Even if the object spans the whole retina, we do not need the highest resolution to classify it, since the resolution will be lost due to the constraint needed for stability. Therefore, matching of the whole object happens only at one level of the multiresolution representation. Each local

feature is first tried for a match at the lowest possible resolution, which is justified by applying the same reasoning. All neurons in the network which perform matching may have a fixed size Receptive Fields (RFs). For the neurons at the lowest resolution it means that the entire input array is covered by a RF (global RFs), and for the neurons at higher resolutions it means that just part of the array is covered (local RFs).

For the purposes of detecting salient structures in the image, we need to have a multiresolution decomposition which differentiates the local orientation of the image features. Such decomposition is an orientation selective wavelet transform [8] or the line segment transform as proposed by Altes [1]. Let us define N wavelet functions $\Psi^i(x,y)(1 \leq i \leq N)$ whose Fourier transform $\hat{\Psi}^i(\omega_x,\omega_y)$ satisfies

$$\sum_{i=1}^{N} \mid \hat{\Psi}^i(\omega_x,\omega_y) \mid^2 = \mid \hat{\Psi}(\omega_x,\omega_y) \mid^2 \tag{2}$$

where $\hat{\Psi}(\omega_x,\omega_y)$ is the Fourier transform of the filter from equation (1). Fig.2 shows an example of decomposition of $\hat{\Psi}(\omega_x,\omega_y)$ into N=8 different wavelets $\hat{\Psi}^i(\omega_x,\omega_y)$.

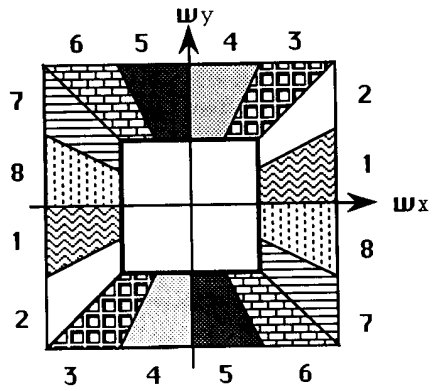

Figure 2: Repartition of the frequency support of orientation selective wavelets at one resolution. Each wavelet is sensitive to one of eight different orientations of local features.

2.2 Detection of Salient Structures

Contemporary neural networks for object recognition can work only with the isolated objects on an empty background. It is apparently not the case with biological visual systems. It is known from neurophysiological findings that the facilitation occurs amongst neurons with cooriented, coaxially aligned receptive fields [12]. Stimulus-related oscillations of neural activities were recently discovered in the primary visual cortex of cat [6] and monkey [5]. These findings led to the hypothesis that synchronization might be a mechanism subserving the transitory linking of local visual features into coherent global percepts.

The detection of salient structures in our model works in the following way. The neurons calculating wavelet coefficients are mutually connected with their neighbors at the same scale of resolution. Neurons having cooriented, coaxially aligned receptive fields mutually facilitate one another, and inhibit the neighbors computing orthogonal orientations. An iterative procedure is applied to find the local winners. The forthcoming implementation will include feature linking via synchronization, like in the model proposed by Eckhorn and coworkers [2].

Since the neurons performing matching have local RFs, they send feedback to the processes detecting salient structures, so that local (under RF) salient structures are preffered. Depending on resolution, these structures will span small areas of an input array (at the highest resolution) or the entire input array (at the lowest resolution). In a sense, the matching neuron enables particular area, and then performs matching of detected structures.

2.3 Object Recognition

Recently there exists a lot of interest in neural networks with locally-tuned neurons [11, 13]. These neurons are more suitable for classification than neurons with sigmoidal response because the local representations ensure that only a few units respond to any given input, thus reducing any computational overhead. As a method of classification, we shall chose some variant of the network with locally-tuned neurons, like the regularization networks.

The line of reasoning in constructing classifier neural networks, advanced by Poggio and coworkers [13], is as follows: (1) for each object there exists a smooth function mapping any perspective view into a "standard" view of the object, and (2) this multivariate function can be synthesized, or at least approximated, from a small number of views of the object. Such a function would be object specific, with different functions corresponding to different objects.

The purpose of synthesizing an approximation to a function from a small number of examples – the views – is to learn an input-output mapping from a set of examples. Applying the regularization theory [13], the following approximation to a function was found:

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i G(\| \mathbf{x} - \mathbf{x}_i \|) \tag{3}$$

where c_i are unknown coefficients, and G is an appropriate basis function, such as the Gaussian function. The views $\mathbf{x}_i$ are the centers for the corresponding basis functions, and $\mathbf{x}$ is an input image.

The coefficients c_i are found during learning by minimizing a measure of the error between the network's prediction and the desired output for each of N examples. Since the problem is not convex, the gradient descent algorithm can get stuck at a local minimum. To overcome this problem we use the ALOPEX algorithm which is proven to avoid local minima[14, 10, 7].

3 Results and Discussion

Computer programs have been written and tested with various simulations that prove the feasibility of the basics of the model, as well as give us directions for future directions.

Fig.3 shows the original input image (top), and the orientation-selective wavelet representation at the highest resolution level (bottom). The original image has 256 x 256 pixels. Each of the eight orientation-selective wavelet representations has 128 x 128 coefficients. The frequency support corresponding to these representations is shown in Fig.2. On the left side of Fig.3 are the first four wavelet representations which correspond respectively to wavelets 1–4 in Fig.2. The four wavelet representations on the right correspond to wavelets 5–8 in Fig.2. These wavelet representations also provide an approximation to the edges in different orientations.

The identification of a particular class member based on local features follows the classification of the object into a proper class. The mechanism for detection of global salient structures directs the entire object classification. The program for detection of salient structures is currently under development. Additional mechanisms for detection of the local salient features direct the focus of attention and allow the comparison of local features. If this bottom-up mechanism fails, the

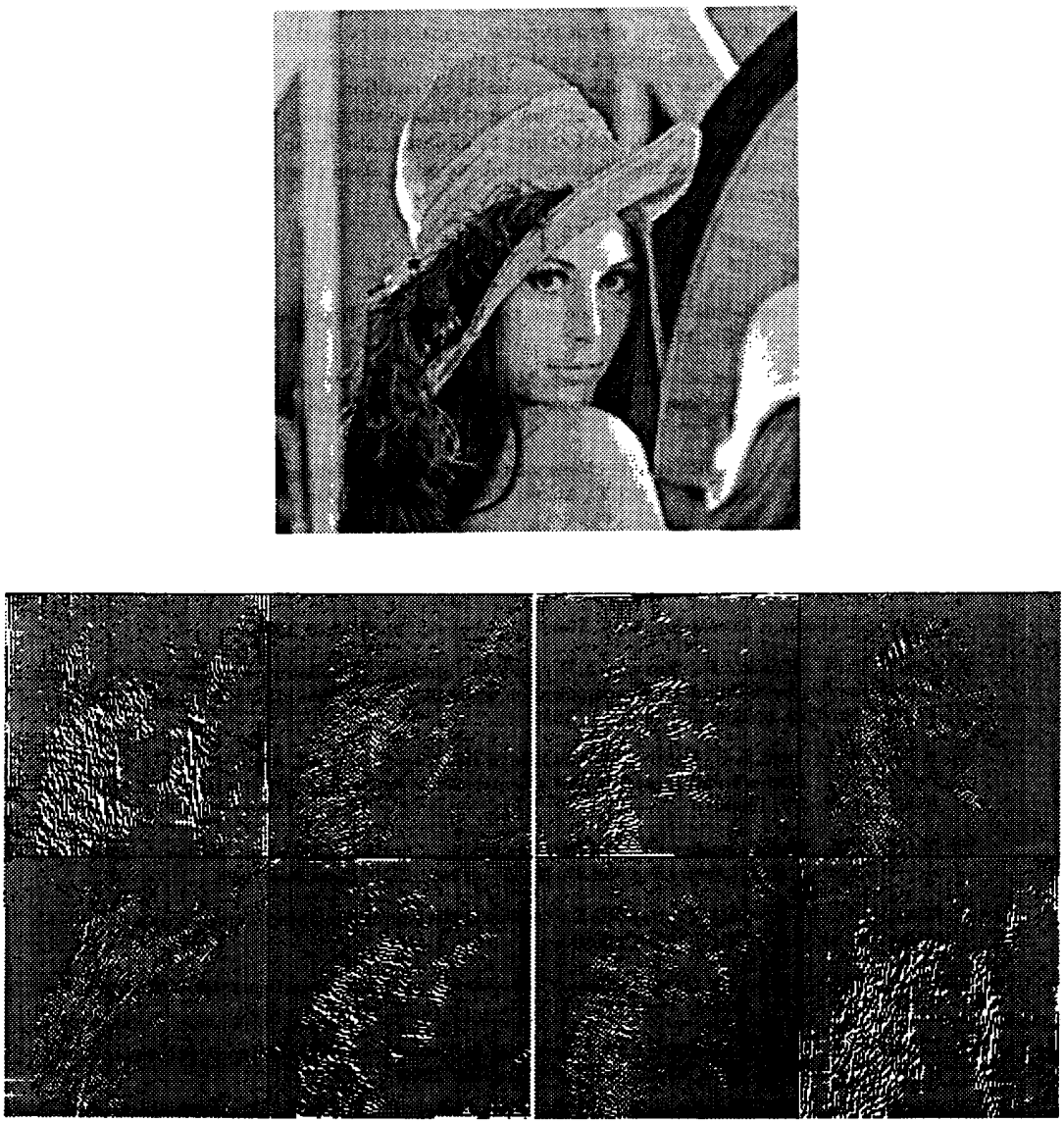

Figure 3: The orientation-selective wavelet representation (bottom) of the original input image (top). Only the highest resolution wavelets are shown. Positive and negative coefficients are shown in light and dark colors respectively. The values close to zero are gray.

top-down processes direct the serial search and the comparison of local features. The whole process takes place within the multiresolution truncated pyramid architecture of neural network [9]. This architecture should allow for easier and more reliable recognition.

Heavily dependent on lateral inhibition and forward excitation, this model shows all characteristics of the biological processes, without trying to duplicate the brain. The current focus of our research is on the detection of salient features, and the integration of global and local mechanisms in the recognition process.

References

[1] ALTES, R.A., "Generalized Wavelet Analysis, the Line Segment Transform, Tomography, and Vision," submitted to *Proc. IEEE*, 1991.

[2] ECKHORN, R., REITBOECK, H.J., ARNDT, M., and P. DICKE, "Feature Linking via Synchronization among Distributed Assemblies: Simulations of Results from Cat Visual Cortex," *Neural Computation*, Vol.2, No.3, pp.293-307, Fall 1990.

[3] EDELMAN, S., and H.H. BÜLTHOFF, "Viewpoint-specific Representations in Three-dimensional Object Recognition," A.I. Memo 1239, M.I.T., Artificial Intelligence Lab, August 1990.

[4] FELLEMAN, D.J., and D.C. VAN ESSEN, "Distributed Hierarchical Processing in Primate Cerebral Cortex: Organization of Macaque Visual Cortex," *Cerebral Cortex*, in press

[5] FREEMAN, W.J., and B.W. VAN DIJK, "Spatial Patterns of Visual Cortical Fast EEG During Conditioned Reflex in a Rhesus Monkey," *Brain Res.*, Vol.422, pp.267-276, 1987.

[6] GRAY, C.M., KÖNIG, P., ENGEL, A.K., and W. SINGER, "Oscillatory Responses in Cat Visual Cortex Exhibit Inter-Columnar Synchronization which Reflects Global Stimulus Properties," *Nature*, Vol.338, No.6213, pp.334-337, 23 March 1989.

[7] HARTH, E.M., and A.S. PANDYA, "Dynamics of the Alopex Process: Application to Optimization Problems," in *Biomathematics and Related Computational Problems*, L. Ricciardi (ed.), Kluwer Acad. Publ., pp.459-471, 1988.

[8] MALLAT, S., "Multifrequency Channel Decompositions of Images and Wavelet Models," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.37, No.12, pp.2091-2110, December 1989.

[9] MARSIC, I., and MICHELI-TZANAKOU, E., "A Neural Mechanism for Object Represenation and Recognition," submitted to NIPS 1991 Conference.

[10] MICHELI-TZANAKOU, E., "When a Feature Detector Becomes a Feature Generator," *IEEE Engineering in Medicine and Biology Magazine*, Vol.9, No.3, pp.19-22, September 1990.

[11] MOODY, J., and C.J. DARKEN, "Fast Learning in Networks of Locally-Tuned Processing Units," *Neural Computation*, Vol.1, No.2, pp.281-294, Summer 1989.

[12] NELSON, J.I., and B.J. FROST, "Intracortical Facilitation among Co-oriented, Co-axially Aligned Simple Cells in Cat Striate Cortex," *Exp. Brain Res.*, Vol.61, No.1, pp.54-61, 1985.

[13] POGGIO, T., and F. GIROSI, "Networks for Approximation and Learning," *Proc. IEEE*, Vol.78, No.9, pp.1481-1497, September 1990.

[14] TZANAKOU, E., MICHALAK, R., and E. HARTH, "The Alopex Process: Visual Receptive Fields by Response Feedback," *Biolog. Cybern.*, Vol.35, pp.161-174, 1979.