A Framework for Object Representation and Recognition[1]

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

Abstract. The objective of this paper is to transform the recognition task into simpler subtasks. Two assumptions are essential in this approach: (a) that the object representation is pictorial, and (b) that the parts of the object do not bear any information about the shape of the object. The second assumption allows the division of the recognition process into classification and identification. Our aim is to find a framework which will make the problem of recognition easier rather than propose another classification neural network. The intractable problem of recognition in high–resolution images is transformed into the problem of classification of iconic representations coupled with the mechanism of focal attention. The set of icons together with their spatial relations is called a pattern, which serves in object identification. The model satisfies the requirement of global computation yet it works with a small number of pixels. It provides a framework in which the contemporary neural networks being applied to simple problems may be applied to real world problems of visual object recognition.

1 Introduction

Since geometric shape usually best characterizes an object, most previous efforts in modeling object recognition have used the shape for object description. The fact that the representation of space is essential has obstructed many efforts to devise a scheme for geometric shape representation and recognition. Modeling the task of shape representation and recognition has been dominated by the requirement that space must be represented. Since both tasks deal with geometry, it seemed natural to assume that both tasks will be solved in the same way and within the same module. The first diversion from this philosophy was the finding by Mishkin and his colleagues that space and shape are represented separately in biological visual systems [13]. Only recently have the modelers of visual object recognition become interested in this fact and begun to exploit it [9, 2].

A common fallacy about the represention of objects is that they are never seen as entities: an object is always thought of as a conglomerate of its parts. In this case, space is unnecessarily introduced into the shape. Because the shape is understood as a conglomerate of parts, we need a geometry to glue them together. Researchers in computer vision have fully accepted the structuralist view and conducted a supposedly simpler search for parts representation in terms of *shape primitives*. Shape primitives are understood as building blocks of shapes [6]. They are usually required to be volumetric (3–D), and each of them (generalized cones, superquadrics, etc.) is actually representing a shape by some simpler shape.

We notice that each part of an object is an object itself. We do not need visual primitives because objects are primitives themselves. Of course, this imposes lower resolution of representation and an inability to distinguish similar objects (classmates), unless we understand that these are distinguished by sub-objects (i.e. objects). We make an important clarification by stating that an object is always recognized in parallel regardless of the complexity of its shape. If an object is described by several objects, i.e. by its parts, then we call it a pattern, and pattern recognition is performed sequentially.

Our assumption is that object representation is pictorial or, equivalently, iconic or unstructured. It may be derived from the fact that biological visual systems have a separate subsystem for spatial vision, which is in agreement with recent psychophysical findings [1].

[1]Appears in: *Proceedings of the International Joint Conference on Neural Networks, Vol.3*, Baltimore, MD, pp.272–277, June 1992.

The inability to differentiate two objects in the same class raises the need for patterns. Noticing the non–uniformities within an object leads to an emergence of new objects inside an object. Once we determine that an object has a part, we append the icon of that part to the existing set of icons. We then create a structure which describes the relationship of this new icon to that of the original object. Adding representations of parts to the set of object representations does not change the parent object representation. It is important to emphasize that the representation of new objects inside an object is equivalent to their representation outside an object. A system which represents these relationships may call these structures scenes and structured objects, respectively, but to avoid confusion we shall refer to them as *patterns* in both cases.

It is important to distinguish icons from patterns. Icons are equivalent to contiguous sets of points in 2–D space. Patterns are equivalent to hierarchical labeled graphs in which vertices are labeled by icons of objects and edges are labeled by their spatial relationships. Each pattern refers to the icons, but it is not itself an icon. This dichotomy of icons and patterns is the basis of the mechanism proposed here. The same part of an image may be represented as an icon or as a pattern; this is most often the source of confusion.

The rationale behind the idea presented here is as follows. We do not want the local details to interfere with the recognition. If the selected part is large with respect to the object, then the information which it bears is already contained in the shape of the entire object. Thus we do not need parts to draw conclusions about an object's shape; the shape information for the object and that of its sub-objects are *dissociated* from one another.

The shape of a sub-object, i.e. a part, does not intrinsically bear any information about the shape of the object. This fact has been extensively exploited in the field of computer vision, where various simple objects are used in object representation (for example spheres, cubes, cylinders, cones, etc.). This is the reason why, in order to recognize the shape of an object, the system does not need to recognize the shapes of its sub-objects. In other words, object recognition initially is distinct from pattern recognition: it is the determination of the object class. Nonetheless, object recognition will eventually include pattern recognition in order to determine a *subclass* or *particular instance* of an object. Another object (part of the original object) is used to distinguish the original object from its classmates. In order to use information about the part in object recognition, the part itself should be recognized as an object.

Thus the recognition consists of two subtasks: *classification* of the object into its proper class and *identification* of the particular member of the class. The classification is performed on the basis of the object's iconic representation; the identification is based on the pattern representation. This fact is used here to propose a multiresolution architecture which features classification of the whole object at only one resolution.

2 Multiresolution Architecture

Recognition is a process which requires global computation. This condition makes it impossible to apply contemporary artificial neural networks to higher resolution images with 128×128 pixels or more. This problem cannot be solved simply by parallel processing or by looking for a better neural network (NN) for classification or neurons with more powerful functions [12, 3]. We must apply some heuristics and look for possible simplifications in the problem itself. We thus propose a mechanism to handle higher resolution images which satisfies the globality requirement yet still works with small numbers of pixels.

A natural way to achieve stability and sensitivity of the recognition process is through a multiresolution representation [6]. Different scales of objects in the image are best described by appropriate spatial frequencies in the Fourier transform of the image. Different frequency bands contain charac-

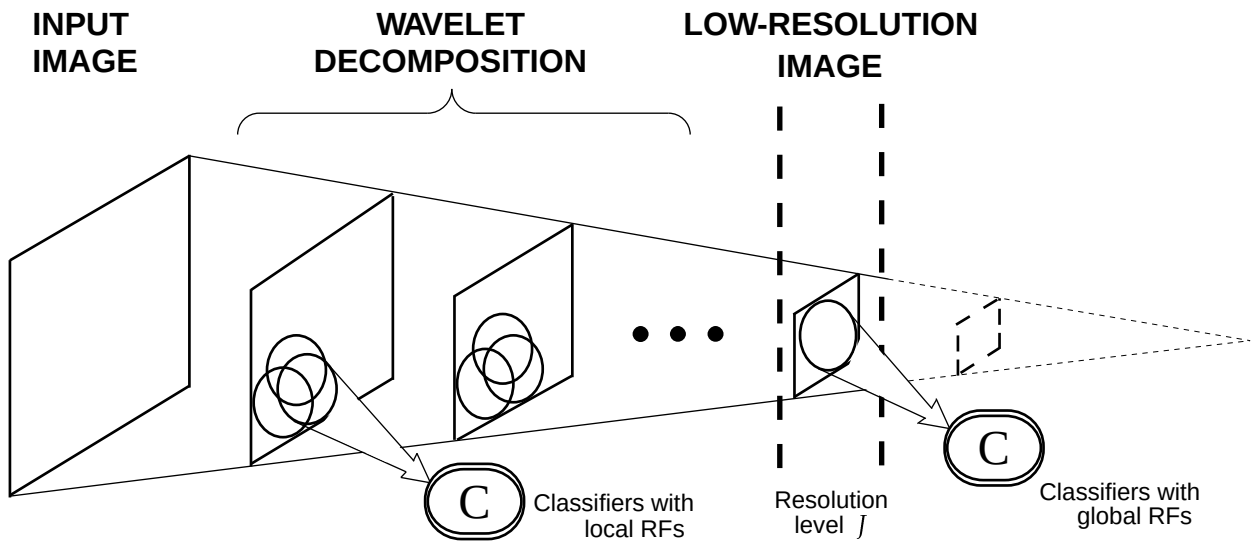

Figure 1: The truncated multiresolution pyramid architecture which allows for stability and sensitivity of the representation. The multiresolution input decomposition is calculated by the wavelet transform. Of the neurons performing object classification, only those at one layer (the lowest resolution) have global RF's. The neurons at higher resolutions have local RF's.

teristic types of information in the image. Low frequencies describe the large objects, spanning the significant portion of an image. Mid frequencies describe their sub-objects, and high frequencies describe fine texture and detail, i.e. objects which are small relative to the size of the entire image. This kind of processing seems to be taking place in biological visual systems, as revealed by psychophysical experiments and neurophysiological recordings[6].

Local details (sub-objects) of an object do not say anything about the object's shape since they belong to different resolution scales. Since details are not important in object classification we have decided to perform classification at the lowest resolution at which it is possible. If this resolution seems too high, it means that we cannot go to a lower one because the object is localized in the entire input array and thus classification may be performed on a small part of the image. The identification or distinction of an object from its classmates is based on details (which are also objects). Otherwise it is already included in the object's description, and it will send the object to a different class. Therefore, identification can also be performed on small parts of the image.

Accordingly, by using a low resolution image and only local portions of higher resolutions, we can recognize objects of any size while satisfying the requirement for globality. Also, we assume that classification precedes identification, i.e. object classification precedes pattern recognition because the pattern refers to icons of objects, and not vice versa.

Thus, at a particular, probably low resolution scale we may classify the whole object, and at a higher resolution we may identify a particular member within the class and represent it as a pattern. This automatically ensures stability and sensitivity of the representation. The low resolution representation partitions the input configuration space into larger subspaces and provides a stable representation. On the other hand, higher resolution and pattern representation provide the possibility to subdivide these subspaces. For both the representation and recognition purposes, it is reasonable to deal with only several higher levels of resolution, not the whole multiresolution pyramid.

Instead of one input configuration, we have several input configurations representing different resolutions. Let us use some classification scheme for recognition at each particular resolution. If the classifier works properly, then classification at any scale where it is possible should produce the

same object class. This assumption is valid because of the requirement for stability, i.e. missing local details should not interfere with the recognition of the whole object. The redundancy of having object classification at several resolution levels may be unnecessary; in addition, it leaves the sensitivity problem yet to be addressed.

We thus propose that *classifying the whole object occurs at only one level of the multiresolution input representation.* Notice that the dissociation of shape information for objects and sub-objects is crucial for this proposition. Let us for a moment leave aside the question of which particular level it should be, and consider the consequences of such a proposition.

The multiresolution input representation may be implemented by any of the well known pyramidal multiresolution techniques, e.g. the Laplacian pyramid or the wavelet pyramid [5]. Let us consider a square image with n^2 sites and employ s as the scaling factor between two resolutions. The image at resolution j will have $(s^{-j}n)^2$ sites. Using the same reasoning as before, we deduce the following characteristics of the multiresolution NN architecture (Fig. 1):

1. Classifiers from only one layer have global receptive fields (RF's) over the entire retina. With the assumption that this is the resolution J, the layer has $(s^{-J}n)^2$ sites.

2. The layer above this one has $(s^{-J+1}n)^2$ sites, and its classifiers have RF's with $(s^{-J}n)^2$ sites.

3. This continues until the last layer, which is the size of the entire input image, and its classifiers also have RF's with $(s^{-J}n)^2$ sites.

Thus, we can apply the same classifiers at all resolutions, since all resolutions have RF's of equal size, namely $(s^{-J}n)^2$ sites. For a classifier at the lowest resolution, this implies that the entire input array is covered by its (global) RF; for any classifier at a higher resolution, it means that just a part of the array is covered by its (local) RF. The choice of the exact RF size for the classifiers at higher resolutions is arbitrary. The topmost level of the truncated multiresolution pyramid, which corresponds to this size, is determined by the desired memory capacity, i.e. how many *object classes* we want to have. Another open question involves the density of classifiers associated with the higher resolution input images, which will be addressed in future communications.

3 Focal Attention

The overall architecture of this recognition system is determined by the truncated multiresolution pyramid derived in the previous section. In this section, we will describe how the processes that perform classification of objects are organized around these input representations.

Although these different resolution input representations are similar, they are calculated by different groups of neurons. This raises the question of how to connect the classification network to them. One possibility is to have only one classification network coupled with a switching mechanism which presents inputs from different resolutions and from different portions of the image. Another possibility is to duplicate the classification network for each resolution and for each portion of the image.

In the first case, the problem consists of controlling the switching mechanism. In the second case, every classification network must be trained separately, and each of their outputs must be the same for the same object inside the receptive field. In other words, we have the problem of assuring the consistency of object representation across different classifiers. This may be accomplished by an additional layer of neurons in the way in which the regularization network generalizes across the different perspective views (see [8])[2].

[2]The mechanisms for small object/part recognition need not be implemented over the whole retina. They may instead be constrained to the central part of the retina; other locations are accounted for by eye movements.

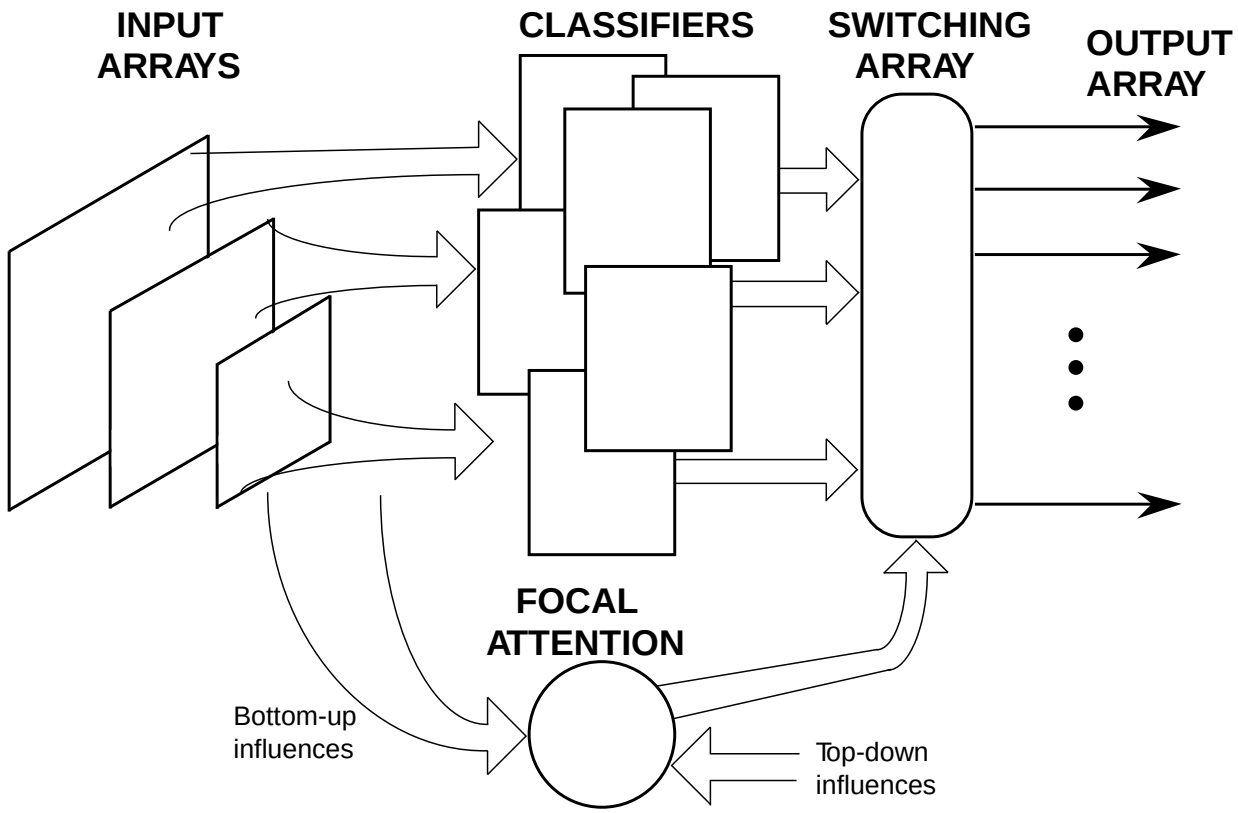

Figure 2: Each classifier classifies an object within its RF. The classifiers do not interact. The focal attention mechanism acts to select a particular classifier which will convey its results to the output array.

Current neurophysiological and psychophysical knowledge is not decisive on this issue. It is very unlikely that the brain implements the physical switching of data streams to one classification device. Findings about the human visual system [7] show that it is not invariant to translations, i.e. an object must be relearned at a new retinal location in order to be recognized. This may favor the case of multiple classifiers, one for each particular location, since the human system will otherwise be completely translation invariant.

On the other hand, findings of Julesz and coworkers [4] about the limited capacity system suggest that the visual system cannot perform recognitions at two different retinal positions simultaneously, i.e. recognition is a serial process. The following solution is plausible in this model. Each portion of every level of the multiresolution input representation has its own associated classifier. The classifiers do not convey their results in parallel, but one at a time. We call this mechanism of allowing the particular classifier to convey its result further the *focal attention mechanism* (Fig. 2), analogous to focal attention in biological vision.

The reason for introducing the (localized) focal attention is the following. The system should classify one object at a time. The entire model of the truncated multiresolution pyramid is derived based on the assumption that other objects, including sub-objects, do not bear any information about the shape of a particular object, i.e. they cannot assist in a bottom–up type of classification. Therefore, only one classifier in the multiresolution pyramid will have the necessary information about a particular object and thus it should be the only one permitted to convey its information to subsequent processes. This applies not only to classifiers at different locations but also to those at the same location but at different resolutions. Therefore, the reason for introducing focal attention in our model is the

interference among outputs of different classifiers, not the limited resources of the visual system.

The most important problem is directing the attention, i.e. selecting a target among the distractors. The target is associated with the classifier which will be allowed to convey its results into the output array; all other structures in the input array are distractors. Focal attention in biological visual systems is controlled both by bottom–up and top–down information [4]. A stimulus will attract the attention if it is particularly salient, if its location has been "precued," or if the subject has been instructed to concentrate on the location where a stimulus may appear. We are primarily interested in control by saliency, i.e. the bottom–up influences, because the others include influences outside the domain of vision.

The saliency of the image regions may be determined in different ways. Examples are texton gradients [4], salient curves characterized by low curvatures [11], etc. There are some specifics of our multiresolution architecture regarding the computation of salient global features. For a classifier in the multiresolution architecture, the input configuration is the portion of the multiresolution input representation spanned by its RF. Features extending outside its RF are irrelevant for a particular classifier. Depending on resolution, salient structures will span small areas of an input array (at the highest resolution) or the entire input array (at the lowest resolution). The detection of salient structures also results in the figure/ground segregation, since the salient structure represents the figure. The matching neuron (or network) then performs classification of detected figures.

4 Classification and Pattern Recognition

The entire analysis up to this point may be seen as a preprocessing stage. The purpose of preprocessing is to isolate the objects in an image and to prepare each one for matching. An object is selected by choosing a classifier within whose receptive field the object is found. The next stage involves matching of the selected object to memorized models, i.e. classification. As a method of classification, we may choose any of the currently available NN's (e.g. [8]).

Another problem is object identification, or pattern recognition, which includes space representation. Space representation faces the same problem as object classification regarding the complexity of learning studied by Judd [3] and therefore it should utilize the same multiresolution architecture (Fig. 1). Besides, in our framework, objects are represented by icons. The locations of sub-objects (parts) are represented within the same 2–D space, which has a one–to–one correspondence with the object's icon's points in the so–called "retinotopic coordinate system." This kind of representation immediately solves the problem of representing the relative size of an object and its sub-object, since the relative size is determined by the distance of the object's and its sub-object's resolution scales. Both relative spatial position and relative size are stored implicitly, although space information needs to be represented explicitely as well.

As a point of particular interest, there exist recent neurophysiological findings by Miyashita and his colleagues [10] about the neurons which may be implementing exactly this kind of pattern representation. Neurons in the anterior temporal cortex of the macaque monkey can be taught to associate pairs of objects with arbitrary shapes. The only similarity among the particular objects selected by different neurons is that they are presented sequentially (or nearly so) during training trials. This suggests that the temporal cortex learns the serial order of the objects, and somehow associates consecutively presented objects with one another.

Since object representation by icons is not rotation invariant and each rotated object produces a new icon, pattern representation needs information about only the part's location, not its orientation.

5 Discussion

Starting from two assumptions, namely (a) the pictorial object representation and (b) the division of the recognition process into classification and identification, we have devised a multiresolution model of object representation and recognition. Instead of proposing another classification NN, we propose a framework which will make the problem of recognition easier. The intractable problem of recognition in high–resolution images is transformed to the problem of recognition in small images coupled with the mechanism of focal attention. It satisfies the requirement of global computation yet it still works with small number of pixels. The model provides a framework in which the neural networks currently being applied to simple problems may be applied to more complex problems of visual object recognition. Besides this practical use, it may also provide insight in explaining some neurophysiological and psychophysical findings.

The purpose of our current research is to reexamine and to further develop the model sketched here. Computer programs to test the model are under development.

References

[1] 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.

[2] Jacobs, R.A., Jordan, M.I., and A.G. Barto, "Task Decomposition Through Competition in a Modular Connectionist Architecture: The What and Where Vision Tasks," *Cognitive Sci.*, Vol.15, pp.219–250, 1991.

[3] Judd, J.S., *Neural Network Design and the Complexity of Learning*, The MIT Press, Cambridge, MA, 1990.

[4] Julesz, B., "Early Vision and Focal Attention", *Rev. Mod. Phys.*, Vol.63, No.3, pp.735–772, July 1991.

[5] 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.

[6] Marr, D.C., *Vision*, W.H. Freeman and Co., New York, 1982.

[7] Nazir, T.A., and J.K. O'Regan, "Some Results on Translation Invariance in the Human Visual System," *Spatial Vision*, Vol.5, No.1, 1990.

[8] Poggio, T., and S. Edelman, "A Network That Learns to Recognize Three–dimensional Objects," *Nature*, Vol.343, pp.263–266, January 1990.

[9] Rueckl, J.G., Cave, K.R., and S.M. Kosslyn, "Why are 'What' and 'Where' Processed by Separate Cortical Visual Systems? A Computational Investigation," *J. Cognitive Neurocsi.*, Vol.1, No.2, pp.171–186, 1989.

[10] Sakai, K., and Y. Miyashita, "Neural Organization for the Long–term Memory of Paired Associates," *Nature*, Vol.354, pp.152–155, November 1991.

[11] Shashua, A., and S. Ullman, "Grouping Contours by Iterated Pairing Network," in *Neural Information Processing Systems 3*, R.P. Lippman, J.E. Moody, and D.S. Touretzky (eds.), Morgan Kauffman Publ., pp.335–341, 1991.

[12] Tsotsos, J.K., "Analyzing Vision at the Complexity Level," *Behavioral Brain Sci.*,Vol.13, pp.423–469, 1990.

[13] Ungerleider, L.G., and M. Mishkin, "Two Cortical Visual Systems," in *Analysis of Visual Behavior*, D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield (eds), The MIT Press, Cambridge, MA, 1982.