MEMO #93–10
July 1993

IDIAP

An RBF Network that Learns Some Aspects of Perceptual Organization

Thomas M. Breuel
IDIAP, C.P. 609, 1920 Martigny, Switzerland
tmb@idiap.ch

Introduction. The "grandmother cell" has a long history in vision. Sometimes it is being proposed as a serious model for recognition problems,[4] and, more often, it is given as a straw man for how recognition could not possibly work.[9] In its usual form, the grandmother cell is described detecting the presence of features in particular locations of an image and firing when enough evidence for a grandmother (or whatever other kind of object the cell is specific for) has been found.

Recently, there has been renewed interest in the grandmother cell because it can be shown that a small collection of such cells, each specific for some 2D view of a 3D object, can collectively approximate well the visual appearance of a 3D object viewed from different angles.[10,1] Those theoretical results together with psychophysical support[5] suggest that a grandmother cell approach, when worked out properly, might account for some aspects of visual object recognition.

Unfortunately, when applied to actual images, such approaches to recognition are empirically prone to false positive detections (misdetections); this phenomenon can be understood in a statistical framework.[2]

The basic reason is that both because of model variation and because of the presence of occlusions in images, matches are never perfect, and a less than maximal activations of the input features to a grandmother cell must still be considered evidence for the presence of an object. But spurious features can conspire to activate enough of the inputs to a grandmother cell to trigger detection, resulting in a false positive. So, there is a tradeoff between the amount of occlusion and model variation that is tolerated and the probability of misdetection.

This effect can be lessened somewhat by making individual receptive fields smaller and thereby reducing the probability that spurious features activate a feature detector, but this means that the grandmother cell may fail to detect instances of an object that are even just slightly deformed.

The traditional approach towards resolving this problem has been to postulate that attentional and segmentational mechanisms precede the recognition stage. Such mechanisms would limit the input to the grandmother cell (or another recognition mechanism) to the features belonging to one object at a time, thereby reducing the probability that features contribute spuriously to the input of the grandmother cell.

In this work, I suggest an alternative, or rather additional, simple mechanism based on higher-order connections at the inputs of the grandmother cell, motivated by theoretical statistical considerations. The relationships between segmentation and the approach presented here will be discussed at the end.

Network Model. Mathematically, a simple grandmother cell is nothing more than a slightly generalized linear discriminant function

$$G = \Theta(\sum \alpha_i F_i + c) \qquad (1)$$

Here, the F_i denote the activations of individual feature detectors, Θ is some global thresholding function, and c is a constant. In particular, in the Radial Basis Function (RBF) approach to 3D object recognition, the F_i are chosen to be the logarithm of Gaussians, and Θ is the exponential function. But we can also interpret the above equation statistically. Let us assume that the feature detectors F_i have binary outputs. Then, the grandmother cell can be used to compute an estimate of the conditional probability that a particular object is present in the image by choosing $\alpha_i = P(\text{object}|F_i)$ and by choosing $\Theta(x) = e^x$, *under the assumption* that the F_i are mutually independent.

In actual images, this independence assumption is violated. This can be shown by actually measuring correlations in images, and it can be argued for on theoretical grounds by considering an idealized world model.[2]

This observation suggests that we consider a more general model of the grandmother cell:

$$G = \Theta(\sum \alpha_i F_i + \beta_{ij} F_i F_j + c) \tag{2}$$

By choosing non-zero β_{ij}, this model can now also take correlations between features into account when assigning quality-of-match measures[*] to the image for the object represented by the grandmother cell.

Experiments. To test the validity of the above ideas, two networks, one corresponding to Equation 1 (*linear network*) and another one corresponding to Equation 2 (*pair network*), were trained on partially occluded images of randomly generated deformations of blob-like objects (positive training examples) and random collections of background features not containing the object (negative training examples). The networks themselves are shown schematically in Figure 1.

As in the literature,[10] the centers of the radial basis functions F_i were chosen from an unoccluded view of the object. The activation of the F_i by training examples is shown in Figure 2.

The weights α_i and β_{ij} were adjusted using the perceptron learning algorithm,[12,8] either with a fixed update rule, or with updates proportional to the inverse of the number of patterns presented to the learning algorithm so far.

The results show a striking difference between the linear network and the pair network in their ability to determine the presence or absence of an object in an image at a given location. For the chosen density of features in the negative training examples and the size of the receptive field of the radial basis functions, the performance of the linear network is only slightly better than chance, whereas the pair network achieves error rates lower than 1%. Representative learning curves are shown in Figure 3.

As a control, these experiments were also carried out at significantly lower densities of features for the negative training examples, in which case both the linear network and the pair network were capable of learning the presence or absence of the object in the image nearly equally reliably. In another set of experiments, in addition to using a low density of features for the negative training examples, some features of the object were more likely to be occluded than others. The interpretation of the parameters α_i in Equation 1 as conditional probabilities would predict that object features that were less likely to be occluded would receive higher weights in the linear network, which was what was found.

It is interesting to look at the kinds of solutions that were found by the networks. The most important observation is that in the pair network, if two feature detectors F_i and F_j have nearby receptive fields, their corresponding term β_{ij} will be significantly greater than zero. The parameters β_{ij} corresponding to feature detectors that are separated widely are near zero or slightly negative (see Figure 4).

The interpretation of this finding is the following. If two features are close to one another, they are both more likely to be affected by an occlusion in the same way than features that are widely separated. Since the receptive fields of the F_i were chosen not to overlap significantly, no such correlations exist for spurious features, neither in the short range, nor in the long range. By giving more weight to those pairs of features that are significantly correlated for positive examples but have no correlation for negative examples, the pair network can achieve its lower error rate.

Discussion. I do not propose this simple network as an alternative to attention and perceptual organization in the human visual system–there is extensive evidence that such phenomena are real–but instead as a complementary mechanism. The implementation in terms of neural hardware seems so simple (it could potentially even be implemented by non-linear processing in dendritic trees) and the advantages of implementing it so straightforward that it would be very surprising if the human visual system did not take advantage of it.

More importantly, however, attention and perceptual organization require some initial representation to operate on (the *initial percept*[11]). Likewise, neural network models of attention, like the ART network,[3] begin attentional selection by comparing an input vector against stored representations. That is, models of recognition involving perceptual organization generally rely on a first step of analysis that has to proceed without the benefit of any attentional selection or prior organization. It is at this stage that the higher-order network methods discussed above could significantly improve the overall performance of the recognition system.

A different class of models for the use of perceptual organization and attention for recognition relies on bottom-up grouping and segmentation processes.[6,13,7] Such methods generally use a fixed measure of how likely it is that different features originate from the same object. Based on such measures, a good (or optimal) global interpretation is found that divides the image into regions belonging to different objects. Such methods also tend to be iterative in nature. In contrast, the method present here is adaptive, even on a per-model basis, it does not require a global interpretation of the image, and it is strictly feed-forward (rather than iterative).

For simplicity's sake, The above discussion was phrased in terms of a "grandmother cell" approach to recognition, but the same principles apply to other kinds of object recognition systems. The approach to reducing the probability of false matches based on higher-order connections described here should prove useful adjunct to methods based on bottom-up segmentation.

References

[1] Thomas M. Breuel. View-Based Recognition. In *MVA '92, IAPR Workshop on Machine Vision Applications*, December 1992.

[2] Thomas M. Breuel. Higher-Order Statistics in Object Recognition. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 1993.

[3] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen. Attentive supervised learning and recognition by an adaptive resonance system. In G. A. Carpenter and S. Grossberg, editors, *Neural Networks for Vision and Image Processing*. MIT Press, 1992.

[4] S. Edelman and T. Poggio. Bringing the grandmother back into the picture: a memory-based view of object recognition . *International Journal of Pattern Recognition and Artificial Intelligence*, vol.6, no.1:37–61, 1992.

[5] Shimon Edelman and Heinrich H. Buelthoff. Viewpoint Specific Repreesentations in 3D Object Recognition. *Proceedings of the National Academy of Science*, 1990.

[6] Stuart Geman and Don Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741, 1984.

[7] Laurent Hérault and Radu Horaud. Feature Grouping and Figure/Ground Discrimination: A Recursive Neural-Network Approach.

In *IJCNN: International Joint Conference on Neural Networks*, pages 2606–2611. IEEE press, 1991.

[8] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Addison Wesley, 1991.

[9] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company, San Francisco, 1982.

[10] Tomaso Poggio and Shimon Edelman. A network that learns to recognize 3D objects. *Nature*, 343(6255):263–266, 1990.

[11] I. Rock. *The Logic of Perception*. MIT Press, Cambridge, MA, USA, 1983.

[12] F. Rosenblatt. *Principles of Neurodynamics*. Spartan Books, New York, 1962.

[13] A. Shashua and S. Ullman. Structural saliency: the detection of globally salient structures using a locally connected network. In *Proceedings of the International Conference on Computer Vision*, pages 321–327, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

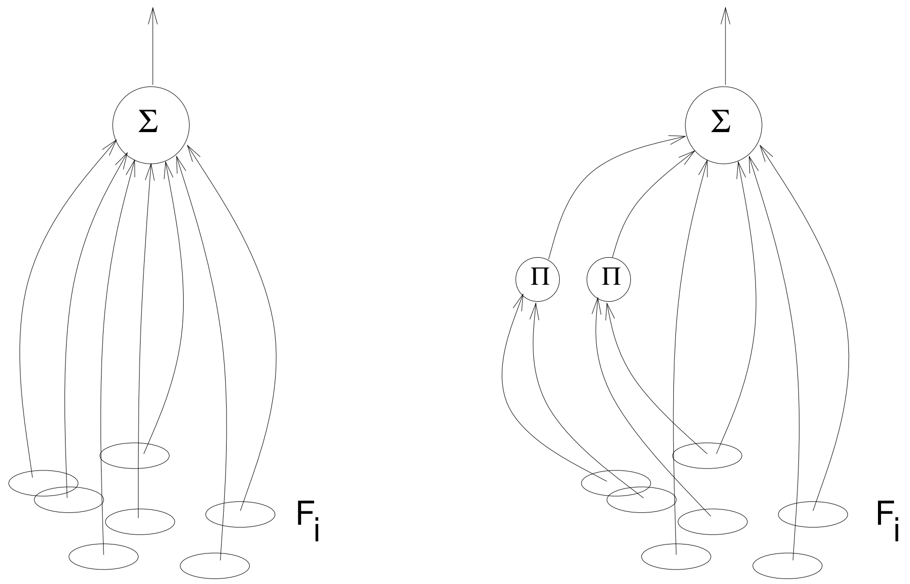

Figure 1: Network Structure.

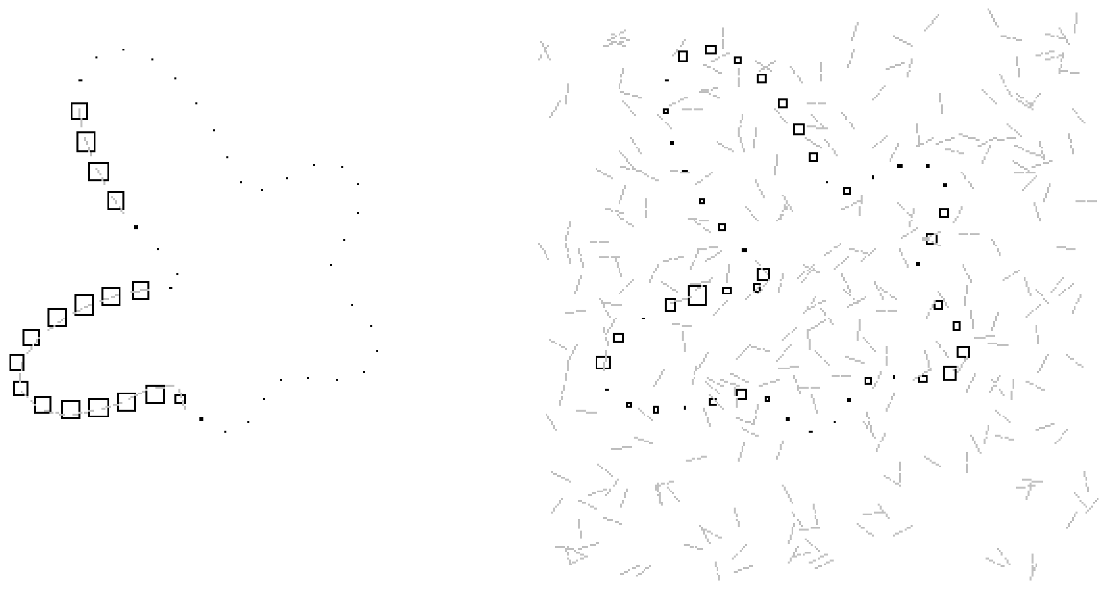

Figure 2: Examples of features (light gray) and activations of features detectors (black squares).

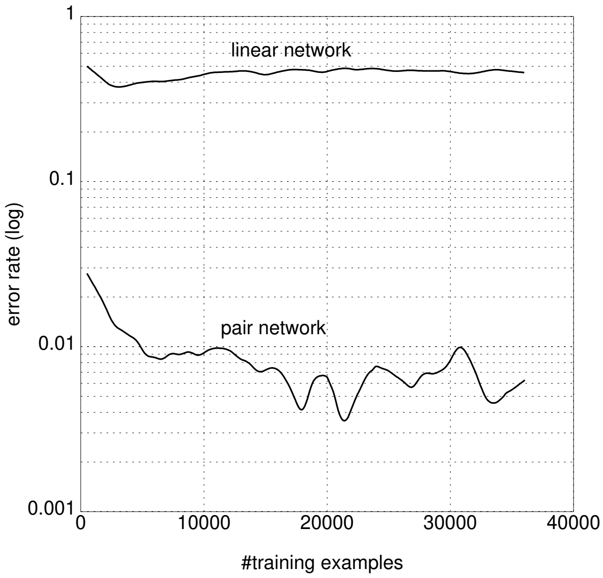

Figure 3: Learning Curves.

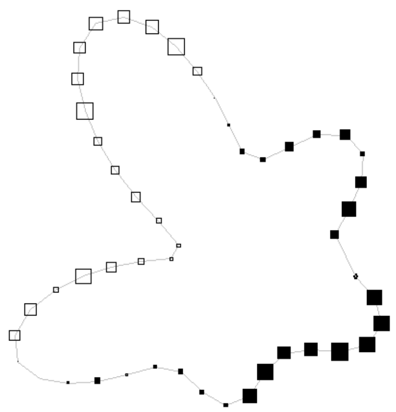

Figure 4: Visualization of the weights β_{0j}. Open squares represent negative weights, filled squares represent positive weights.