MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING

A.I. Memo No.1404
C.B.C.L. Paper No. 77

December 1993

Observations on Cortical Mechanisms for Object Recognition and Learning

Tomaso Poggio and Anya Hurlbert

Abstract

This paper sketches several aspects of a hypothetical cortical architecture for visual object recognition, based on a recent computational model. The scheme relies on modules for learning from examples, such as Hyperbf-like networks, as its basic components. Such models are not intended to be precise theories of the biological circuitry but rather to capture a class of explanations we call Memory-Based Models (MBM) that contains sparse population coding, memory-based recognition and codebooks of prototypes. Unlike the sigmoidal units of some artificial neural networks, the units of MBMs are consistent with the usual description of cortical neurons as tuned to multidimensional optimal stimuli. We will describe how an example of MBM may be realized in terms of cortical circuitry and biophysical mechanisms, consistent with psychophysical and physiological data. A number of predictions, testable with physiological techniques, are made.

This memo describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by grants from the Office of Naval Research under contracts N00014-92-J-1879 and N00014-93-1-0385; and by a grant from the National Science Foundation under contract ASC-9217041 (this award includes funds from ARPA provided under the HPCC program). Additional support is provided by the North Atlantic Treaty Organization, ATR Audio and Visual Perception Research Laboratories, Mitsubishi Electric Corporation, Sumitomo Metal Industries, and Siemens AG. Support for the A.I. Laboratory's artificial intelligence research is provided by ARPA contract N00014-91-J-4038. Tomaso Poggio is supported by the Uncas and Helen Whitaker Chair at MIT's Whitaker College.

1 Introduction

One of the main goals of vision is object recognition. But there may be many distinct routes to this goal and the goal itself may come in several forms. Anyone who has struggled to identify a particular amoeba swimming on a microscope slide or to distinguish between novel visual stimuli in a psychophysics laboratory might admit that recognizing a familiar face seems an altogether different and simpler task. Recent evidence from several lines of research strongly suggests that not all recognition tasks are the same. Psychophysical results and computational analyses suggest that recognition strategies may depend on the type of both object and visual task. Symmetric objects are better recognized from novel viewpoints than asymmetric objects (Poggio and Vetter, 1992); when moved to novel locations in the visual field, objects with translation-invariant features are better recognized than those without (Bricolo and Bülthoff, 1992; Nazir and O'Regan, 1990). A typical agnosic patient can distinguish between a face and a car, a classification task at the *basic* level of recognition, but cannot recognize the face of Marilyn Monroe, an identification task at the *subordinate* level (Damasio and Tranel, 1990). A recently-reported stroke patient cannot identify the orientation of a line but can align her hand with it if she imagines posting a letter through it, suggesting strongly that there are also multiple outputs from visual recognition (Goodale, 1991).

Yet although recognition strategies diverge, recent theories of object recognition converge on one mechanism that might underlie several of the distinct stages, as we will argue in this paper. This mechanism is a simple one, closely related to template matching and Nearest Neighbor techniques. It belongs to a class of explanations that we call Memory-Based Models (MBMs), which includes memory-based recognition, sparse population coding, Generalized Radial Basis Functions networks, and their extension, Hyper Basis Functions networks (HBF) (Poggio and Girosi, 1990b) (see Figure 2.) In MBMs, classification or identification of a visual stimulus is accomplished by a network of units. Each unit is broadly tuned to a particular template, so that it is maximally excited when the stimulus exactly matches its template but also responds proportionately less to similar stimuli. The weighted sum of activities of all the units uniquely labels a novel stimulus. Several recent and successful face recognition schemes for machine vision share aspects of this framework (Baron, 1981; Bichsel, 1991; Brunelli and Poggio, 1992; Turk and Pentland, 1991; Stringa, 1992a; Stringa, 1992b)

We will consider how the basic features of this class of models might be implemented by the human visual system. Our aim is to demonstrate that such models conform to existing physiological data and to make further physiological predictions. We will use as a specific example of the class the RBF network. RBF networks /footnoteWe use the term RBF here in a broad sense including generalizations of the pure RBF scheme such as GRB and HBF (see Poggio and Girosi 1990). have been used successfully to solve isolated visual tasks, such as learning to detect displacements at hyperacuity resolution (Poggio, Fahle and Edelman, 1992) or learning to identify the gender of a face (Brunelli and Poggio, 1992). We will discuss how the units of a RBF network might be realized as neurons and how a similar network might be implemented by cortical circuitry and replicated at many levels to perform the multi-component task of visual recognition. We hope to show that MBMs are not merely toy replicas of neural systems, but viable models that make testable biological predictions.

The main predictions of Memory-Based Models are:

- The existence of broadly tuned neurons at all levels of the visual pathway, tuned to single features or to configurations in a multidimensional feature space.
- At least two types of plasticity in the adult brain, corresponding to two stages of learning in perceptual skills and tasks. One stage probably involves changes in the tuning of individual neuron responses; this resembles adaptation. The other probably requires changes in cortical circuitry specific to the task being learned, connecting many neurons across possibly many areas.

2 Object Recognition: Multiple Tasks, Multiple Pathways

Recognizing an object should be difficult because it rarely looks the same on each sighting. Consider the prototypical problem of recognizing a specific face. (We believe that processing of faces is not qualitatively different from processing other 3D objects, although the former might be streamlined by practice, and biological evidence supports this view [Gross, 1992].) The 2D retinal image formed by the face changes with the observer's viewpoint, and with the many transformations that the face can undergo: changes in its location, pose, and illumination, as well as non-rigid deformations such as the transition from a smile to a frown. A successful recognition system must be robust under all such transformations.

Here we outline an architecture for a recognition system that contains what we believe are the rudimentary elements of a robust system. It is best considered as a protocol for and summary of existing programs in machine vision, but it also represents an attempt to delineate the stages probably involved in visual recognition by humans. The scheme (diagrammed in Figure 1) has dual routes to recognition. The first is a streamlined route to recognition in which the features extracted in the early stages of image analysis are matched directly to samples in the database. The second potential route to recognition diverges from the first to allow for the possibility that both the database models and the extracted image features might need further processing before a match can be found.

Our task in recognizing a face – or any other 3D object – consists of multiple tasks, which fall into three broad categories that characterize both routes:

- *Segmentation*: Marking the boundaries of the face in the image. This stage typically involves segmenting the entire image into regions *likely* to correspond to different materials or surfaces (and thereby subsumes figure-ground segmentation) and is a prerequisite for further analysis of a marked region. *Image measurements* are used to convert the retinal array of light intensities into a *primal image representation*, by computing sparse measurements on the array, such as intensity gradients, or center-surround outputs. The result is a set of vector measurements at each of a sparse or dense set of locations in the image. These measurements may be global ones like average value over a whole array of (filtered) pixel values.
- *Classification*, or basic-level recognition: Distinguishing objects that are faces from those that are not. Parameter values estimated in the preceding stage – e.g. the distance between eyes and mouth – are used in this stage for classification of a set of features – e.g. as a potential face, animal, or man-made tool. This stage requires that the boundaries or the location of at least potential faces be demarcated, and hence generally depends on the preceding step of image segmentation, although it may work without it at an added computational cost.
- *Identification*, or subordinate-level recognition: Matching the face to a stored memory, and thereby labeling it. This stage requires some form of indexing of the database samples. Because it is computationally implausible that the recognition system contains a stored sample of the face in each of its possible views or expressions, or under all possible illumination conditions at all possible viewing distances, this step in general also requires that the face be transformed into a standard form for matching against its stored template. Thus in parallel with the direct route from classification to identification there may exist a second route that we call the *visualization* route, which may include an iterative sequence of transformations of both the image-plane and the database models until it converges on a match.

These stages, and some open questions on the overall architecture, are further discussed in the Appendices.

As outlined here, the stages are distinct and could be implemented in series within each route to recognition. Most artificial face recognition systems tackle the stages separately, being designed either to detect and localize a face in an image cluttered with other objects (segmentation and classification), or to identify individual faces presented in an expected format (database indexing and identification). Some artificial recognition systems have been constructed to achieve invariant recognition under isolated transformations (visualization). Examples are systems that: recognize frontal views of faces under varying illuminations (Brunelli and Poggio, 1992); recognize simple paper-clip-like objects independently of viewpoint (Poggio and Edelman, 1990); or identify simple objects solely by color under spatially varying illumination (Swain and Ballard, 1990).

Yet in biological systems, and in some artificial systems, the stages may act in parallel or even merge. For example, there may be many short-cuts to recognizing a frequently encountered object such as a face, for example.

Finding the face might be streamlined by a quick search at low resolution over the whole image for face-like patterns. The search might employ simplified templates of a face containing anthropometric information (for example, a two-eyes-and-mouth mask). Once located, salient features such as eyes can be used to demarcate the entire object to which they belong, eliminating the need to segment other parts of the image. These detectors would scan the image for the presence of these face-specific features, and using them, locate the face for further processing (translation, scaling, etc.). (Some machine vision systems already implement this idea, using translation-invariant face-feature-detectors such as eye detectors [Bichsel, 1991] or symmetry detectors.) Thus segmentation may occur simultaneously with classification. The existence of these face-detectors in the human visual system might explain why we so readily perceive faces in the simplest drawings of dots and lines, or in symmetric patterns formed in nature (Hurlbert and Poggio, 1986), and why we detect properly configured faces more readily than arbitrary or inverted arrangements of facial features (Purcell and Stewart, 1988). Indeed, we wonder whether face recognition may have become so inveterate that the human brain might first classify image regions into face or non-face. Notice that the process of finding features such as the eyes and identifying the face are probably very similar in this view. They are both based on a set of prototypical examples of either eyes or views of the particular face, and they may be using a similar machinery perhaps (RBF-like).

Recognizing an expected object might also be more speedy and efficient than identifying an unexpected one. In the classification stage, only those features specific for the expected object class need be measured, and correct classification would not require that all features be simultaneously available. This step might therefore be itself a form of template matching, where part-templates may serve as well as whole-templates to locate and classify the object. In many cases the classification stage may lead by itself to unique recognition, especially when situational information, such as the expectedness of the object, restricts the relevant data base.

Yet many questions are left hanging by this sketch of a recognition system. In biological systems, is matching done between primal image representations, like center-surround outputs at *sparse* locations, or between sets of higher level features? Computational experiments on face recognition suggest that the former strategy performs much better. What exactly are the key features used for identifying, localizing and normalizing an object of a specific class? Is there an automatic way to learn them? (Huber, 1985). Do biological visual systems acquire recognition features through experience (Edelman, 1991)? Do humans use expectation to restrict the data base for categorization? Some psychophysical experiments suggest that we do not need higher-level expectations to recognize objects quickly in a random series of images, but these experiments have used familiar objects such as the Eiffel Tower (M. Potter, pers. comm.).

2.1 A Sketch of a Memory-Based Cortical Architecture for Recognition

We suggest that most stages in face recognition, and more generally, in object recognition, may be implemented by modules with the same intrinsic structure – a Memory Based Module (MBM). At the heart of this structure is a set of neurons each tuned to a particular value or configuration along one or many feature dimensions. Let us take as an example of such a structure the Hyper Basis Functions (HBF) network. This is a convenient choice because HBFs have been successfully applied already to several problems in object recognition as well as an unrestrictive, easily modifiable choice because HBFs are closely related to other approximation and learning techniques such as multilayer perceptrons.

2.1.1 RBF Networks

HBF networks are approximation schemes based on, but more flexible than, Radial Basis Functions (RBF) networks (see Figure 2; Poggio and Girosi, 1990b; Poggio, 1990). The fundamental equation underlying RBF networks states that any function $f(\mathbf{x})$ may be approximated by a weighted sum of RBFs:

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i h(||\mathbf{x} - \mathbf{t_i}||)^2 + p(\mathbf{x}). \qquad (1)$$

The functions h may be any of the class of RBFs, for example, Gaussians. $p(\mathbf{x})$ is a polynomial that is required by certain RBFs for the validity of the equation. (For some RBFs, e.g. Gaussians, the addition of $p(\mathbf{x})$ is not necessary, but improves performance of the network.) In an RBF network, each "unit" computes the distance $||\mathbf{x} - \mathbf{t}||$ of the input vector $\mathbf{x}$ from its center $\mathbf{t}$ and applies the function h to the distance value, i.e. it computes the function $h(||\mathbf{x} - \mathbf{t}||)^2$. The N centers $\mathbf{t}$, corresponding to the N data points, thus behave like *templates*, to which the inputs are compared for similarity.

A typical and illustrative choice of RBF is the Gaussian $[h(||\mathbf{x}-\mathbf{t}||) = exp(-(||\mathbf{x}-\mathbf{t}||)^2/2\sigma^2)]$. In the limiting case where h is a very narrow Gaussian, the network effectively becomes a *look-up* table, in which a unit gives a non-zero signal only if the input exactly matches its center $\mathbf{t}$.

The simplest recognition scheme based on RBF networks that we consider is that suggested by Poggio and Edelman (1990) (see fig. 7) to solve the specific problem of recognizing a particular 3D object from novel views, a subordinate-level task. In the RBF version of the network, each center stores a sample view of the object, and acts as a unit with a Gaussian-like recognition field around that view. The unit performs an operation that could be described as "blurred" template matching. At the output of the network the activities of the various units are combined with appropriate weights, found during the learning stage. An example of a recognition field measured psychophysically for an asymmetric object after training with a single view is shown in fig 5. As predicted from the model (see Poggio and Edelman, 1990), the shape of the surface of the recognition errors is roughly Gaussian and centered on the training view.

In this particular model, the inputs to the network are spatial coordinates or measurements of features (e.g. angles or lengths of segments) computed from the image. In general, though, the inputs to an RBF network are not restricted to spatial coordinates but could include, for example, colours or configurations of segments, binocular disparities of features, or texture descriptions. Certainly in any biological implementation of such a network the inputs may include measurements or descriptions of any attribute that the visual system may represent. We assume that in the primate visual system such a recognition module may use a large number of primitive measurements as inputs, taken by different "filters" that can be regarded as many different "templates" for shape, texture, color and so forth. The only restriction is that the features must be directly computed from the image. Hence the inputs are viewer-centered, not object-centered, although some, like colour, will be viewpoint-independent. The output of the network is, though, object-centered, provided there is a sufficient number of centers. This generality of the network permits a mix of 2D and 3D information in the inputs, and relieves the model from the constraints of either.

This feature of the model also renders irrelevant the question on whether object representations are 2D or 3D. The Poggio-Edelman model makes it clear that 2D-based schemes can provide view invariance as readily as a 3D model can, and compute 3D pose as well (see Poggio and Edelman, 1990). So the relevant questions are: what is *explicit* in neurons? and what does it mean for information about shape to be explicit in neurons? In a sense, some 2D-based schemes such as the Poggio-Edelman model may be considered as plausible neurophysiological implementations of 3D models.

We do not suggest that the cortical architecture for recognition consists of a collection of such modules, one for each recognizable object. Certainly it is more complex than that cartoon, and not only because viewpoint-invariance is not the only problem that the recognition system must solve. For example, the cortex must also learn to recognize objects under varying illumination (photometric invariance) and to recognize objects at the basic as well as subordinate level. [Preliminary results on real objects (faces) suggest that HBF modules can estimate expression and direction of illumination equally as well as pose (Brunelli, pers. comm., Beymer, pers. comm.).] Yet each of these and other distinct tasks in recognition may be implemented by a module broadly similar to the Poggio-Edelman viewpoint-invariance network. We might expect that the system could be decomposed into elementary modules similar in design but different in purpose, some specific for individual objects (and therefore solving a subordinate-level task), some specific to an object class (solving a basic-level task), and others designed to perform transformations or feature extractions, for example, common to several classes. The modules may broadly be categorized as:

- *Object-specific.* A module designed to compensate for specific transformations that a specific object might undergo. As in the Poggio-Edelman network, the module would consist of a few units, each maximally tuned to a particular configuration of the object – for the face, say, a particular combination of pose and expression. A more general form of the network may be able to recognize a few different faces: its hidden units would be tuned to different views but of not just one face, and therefore behave more like eigenfaces.

- *Class-specific.* A module that generalizes across objects of a given class. For example, the network may be designed to extract a feature or aspect of *any* of a class of objects, such as pose, color, or distance. For example, there might be a network designed to extract the pose of a face, and a separate network designed to extract the direction of illumination on it. Any face fed as input to network would elicit an estimate of its pose or illumination.

- *Task-specific.* Networks that solve tasks, such as shape-from-shading, *across* classes of objects. An example would be a generic shape-from-shading network that takes as input brightness gradients of image regions. It may act in the early stages of recognition, helping to segment and classify 3D shapes even before they are grouped and classified as objects.

The distinctions between these types of recognition module might be blurred if, for example, the visual system overlearns certain objects or transformations. For example, a shape-from-shading network might develop for a frequently-encountered type of material, or for a specific class of object. Indeed, our working assumption is that any apparent differences between recognition strategies for different types of objects arise not from fundamental differences in cortical mechanisms but from imbalances in the distribution of the same basic modules across different objects and different environments. Savanna Man, like us, probably had task-specific modules dedicated to faces, but although we might have shape-from-shading modules specific to familiar pieces of office furniture, he might not be able to recognize a filing cabinet at all, much less under varying illumination. This suggests a decomposition into modules that are both task and object specific, which is a rather unconventional but plausible idea.

Transformations specific to a particular object may also be generalized from transformations learned on prototypes of the same class. For example, the deformation caused by a change in pose or, for a face, a change in expression or age, may be learned from a set of examples of the same transformation acting on prototypes of the class. Some transformations may be generalized across all objects sharing the same symmetries (Poggio and Vetter, 1992).

The big question is, if the recognition system does consist of similar modules performing interlocking tasks, how are the modules linked, and in what hierarchy (if it makes sense at all to talk of ordered stages)? In constructing a practical system for face recognition, it would make sense first to estimate the pose, expression, and illumination for a generic face and then to use this estimate to "normalize" the face and compare it to single views in the data base (additional "search" to fine tune the match may be necessary). Thus the system would first employ a class-specific module based on invariant properties of faces to recover, say, a generic view – analogous to an object-centered representation – that could feed into face-specific networks for identification. The information that the system extracts in the early stages concerning illumination, expression, context, etc. would not be discarded. Within each stage, modules may be further decomposed and arranged in hierarchies: one may be specific for eyes, and may extract gaze angle, a parameter that may then feed into a module concerned with the pose of the entire face.

For face recognition, the generic view may be recovered by exploiting prior information such as the approximate bilateral symmetry of faces. In general a single monocular view of a 3D object (if shading is neglected) does not contain sufficient 3D information for recognition of novel views. Yet humans are certainly able to recognize faces rotated 20-30 degrees away from frontal after training on just one frontal view. One of us has recently discussed (Poggio, 1991) different ways for solving the following problem: *from one 2D view of a 3D object generate other views, exploiting knowledge of views of other, "prototypical" objects of the same class.* It can be shown theoretically (Poggio and Vetter, 1992) that prior information on generic shape constraints does reduce the amount of information needed to recognize a 3D object, since additional virtual views can be generated from given model views by the appropriate symmetry transformations. In particular, for bilaterally symmetric objects, a single non-accidental "model" view is theoretically sufficient for recognition of novel views. Psychophysical experiments (Vetter, Poggio and Bülthoff, 1992) confirm that humans are better in recognizing symmetric than non-symmetric objects.

An interesting question is whether there are indeed multiple routes to recognition. It is obvious that some of the logically distinct steps in recognition of Figure 1 may be integrated in fewer modules, depending on the specific implementation. Figure 3 shows how the same architecture may appear if the classification and the visualization routes are implemented with HBF networks. In this case, the database of face models would essentially be embedded in the networks (see Poggio and Edelman, 1990).

There are of course several obvious alternatives to this architecture and many possible refinements and extensions. Even if oversimplified, this token architecture is useful to generate meaningful questions. The preceding discussion may in fact be sufficient for performing computational experiments and for developing practical systems. It is also sufficient for suggesting psychophysical experiments. It is of course not enough from the point of view of a physiologist, yet the physiological data in the next section provides broad support for its ingredients.

2.1.2 Physiological Support for a Memory-Based Recognition Architecture

At least superficially, physiological data seems to support the existence of elements of each these modules. Perrett et.al. (Perrett et. al., 1989; Perrett et.al., 1985) report evidence from inferotemporal cortex (IT) not only for cells tuned to individual faces but also for face cells tuned to intermediate views between frontal and profile, units that one would expect in a class-specific network designed to extract pose of faces. Such cells also support the existence of the view-centered units predicted by the basic Poggio-Edelman recognition module. Young and Yamane (1992) describe cells in anterior IT that respond optimally to particular configurations of facial features, or "physical prototypes." These may conceivably provide input to the cells described by Perrett et. al. as "person recognition units", or to the approximately view-independent cells described by Hasselmo, et. al. (Hasselmo et.al., 1989) which would in turn correspond almost exactly to the object-centred output of the Poggio-Edelman model. Perrett et. al. (1989; 1985) also report cells that respond to a given *pose* of the face regardless of illumination – even when the face is under heavy shadow. Such cells may resemble units in a *task-specific* network. In the superior temporal sulcus, Hasselmo et. al. (1989) also find cells sensitive to head movement or facial gesture, independent of the view or identity of the face. Such cells would also appear to be both *class-* and *task-specific*. (See Perrett and Oram, 1992) for a more detailed review of relevant physiological data.)

Fujita and Tanaka (1992) have also reported cells in IT that respond optimally to certain configurations of color and shape. These may well represent elements of networks that generalize across objects, classifying them according to their geometric and material constitution. More significantly, Fujita and Tanaka (1992) report that cells in the anterior region of IT (cytoarchitectonic area TE) are arranged in columns, within which cells respond to similar configurations of color, shape and texture. Each configuration may be thought of as a template, which in turn might encode an entire object (e.g. a face) or a part of an object (e.g. the lips). Within one column, cells may respond to slightly different versions of the template, obtained by rotations in the image-plane, for example. Fujita and Tanaka (1992) conclude that each of the 2000 or so columns in TE may represent one phoneme in the language of objects, and that combinations of activity across the columns are sufficient to encode all recognizable objects.

The existence of such columns supports the notion that the visual system may achieve invariance to image-plane transformations of elementary features by replicating the necessary feature measurements at different positions, at different scales and with different rotations.

In the next section we describe how key aspects of the architecture could be implemented in terms of plausible biophysical mechanisms and neurophysiological circuitries.

3 Neural modeling of memory-based architectures for recognition

In this section we discuss in more detail the possible neural implementations of a recognition system built from MBMs. The main questions we address are: how are MBMs constructed when a new object or class of objects is learned? and how might MBM units be constructed from known biophysical mechanisms? We propose that there are two stages of learning – supervised and unsupervised – and illustrate to which elements of a memory-based network they correspond. Where could they be localized in terms of cortical structures? What mechanisms could be responsible? We discuss the memory-based module itself and the circuitry that might underlie it.

3.1 The learning-from-examples module

The simple RBF version of an MBM, discussed in section 2.1, learns to recognize an object in a straightforward way. Its centers are fixed, chosen as a subset of the training examples. The only parameters that can be modified as the network learns to associate each view with the correct response ("yes" or "no" to the target object) are the coefficients c_i, the weights on the connections from each center to the output.

The full HBF network permits learning mechanisms that are more biologically plausible by allowing more parameters to be modified. HBF networks are equivalent to the following scheme for approximating a multivariate function:

$$f^*(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(\|(\mathbf{x}-\mathbf{t}_\alpha)\|_W^2) + p(\mathbf{x}) \qquad (2)$$

where the centers $\mathbf{t}_\alpha$ and coefficients c_α are unknown, and are in general fewer in number than the data points ($n \leq N$). The norm is a *weighted norm*

$$\|(\mathbf{x}-\mathbf{t}_\alpha)\|_W^2 = (\mathbf{x}-\mathbf{t}_\alpha)^T W^T W(\mathbf{x}-\mathbf{t}_\alpha) \qquad (3)$$

where $\mathbf{W}$ is an unknown square matrix and the superscript T indicates the transpose. In the simple case of diagonal $\mathbf{W}$ the diagonal elements w_i assign a specific weight to each input coordinate, determining in fact the units of measure and the importance of each feature (the matrix $\mathbf{W}$ is especially important in cases in which the input features are of a different type and their relative importance is unknown) (Poggio and Girosi, 1990a). During learning, not only the coefficients c but also the centers $\mathbf{t}_\alpha$, and the elements of $\mathbf{W}$ are updated by instruction on the input-output examples. See Figure 4.

Whereas the RBF technique is similar to and similarly limited as template matching, HBF networks perform a generalization of template matching in an appropriately linearly transformed space, with the appropriate metric. HBF networks are therefore different in both interpretation and capabilities from "vanilla" RBF. An RBF network can recognize an object rotated to novel orientations only if it has centers corresponding to sample rotations of the object. HBFs, though, can perform a variety of more sophisticated recognition tasks. For example, HBFs can:

1. discover the Basri-Ullman result (Basri and Ullman, 1989; Brunelli and Poggio, unpublished). (In its strong form (see Poggio 1991), this result states that under orthographic projection any view of the visible features of the 3D object may be generated by a linear combination of 2 other views.);
2. with a non-diagonal $\mathbf{W}$, recognize an object under orthographic projection with only one center;
3. provide invariance (or near invariance under perspective projection) for scale, rotation and other uniform deformations in the image plane, without requiring that the features be invariant;
4. discover symmetry, collinearity and other "linear-class" properties (see Poggio and Vetter, 1992).

3.1.1 Gaussian Radial Basis Functions

In the special case where the network basis functions are Gaussian and the matrix $\mathbf{W}$ diagonal, its elements w_i have an appealingly obvious interpretation. A multidimensional Gaussian basis function is the product of one-dimensional Gaussians and the scale of each is given by the inverse of w_i. For example, a 2D Gaussian radial function centered on $\mathbf{t}$ can be written as:

$$G(\|\mathbf{x}-\mathbf{t}\|_W^2) \equiv e^{-\|\mathbf{x}-\mathbf{t}\|_W^2} = e^{-\frac{(x-t_x)^2}{2\sigma_x^2}} e^{-\frac{(y-t_y)^2}{2\sigma_y^2}} , \qquad (4)$$

where $\sigma_x = 1/w_1$ and $\sigma_y = 1/w_2$, and w_1 and w_2 are the elements of the diagonal matrix $\mathbf{W}$.

Thus a multidimensional center can be factored in terms of one-dimensional centers. Each one-dimensional center is individually tuned to its input: centers with small w_i, or large σ_i, are less selective and will give appreciable responses to a range of values of the input feature; centers with large w_i, or small σ_i, are more selective for their input and accordingly have greater influence on the response of the multidimensional center. The template represented by the multidimensional center can be considered as a conjunction of one-dimensional templates. In this sense, a Gaussian HBF network performs the disjunction of conjunctions: the conjunctions represented by the multidimensional centers are "or"ed in the weighted sum of center activities that forms the output of the network.

3.2 Expected physiological properties of MBM units

3.2.1 The neurophysiological interpretation of HBF centers

Our key claim is that HBF centers and tuned cortical neurons behave alike.

A Gaussian HBF unit is maximally excited when each component of the input exactly matches each component of the center. Thus the unit is optimally tuned to the stimulus value specified by its center. Units with multidimensional centers are tuned to complex features, formed by the conjunction of simpler features, as described in the previous section.

This description is very like the customary description of cortical cells optimally tuned to a more or less complex stimulus. So-called place coding is the simplest and most universal example of tuning: cells with roughly Gaussian receptive fields have peak sensitivities to given locations in the input space; by overlapping, the cell sensitivities cover all of that space. In V1 the input space may be up to 5 dimensional, depending on whether the cell is tuned not only to the retinal coordinates x, y but also to stimulus orientation, motion direction and binocular disparity. In V4 some cells respond optimally to a stimulus combining the appropriate values of speed and color (N. K. Logothetis, pers. comm.; Logothetis and Charles, 1990). Other V4 cells respond optimally to a combination of colour and shape (D. Van Essen, pers. comm.) . In MST cells exist optimally tuned to specific motions in different parts of the receptive field and therefore to different motion "dimensions". Most of these cells are also selective for stimulus contrast. In "later" areas such as IT cells may be tuned to more complex stimuli which can be changed in a number of "dimensions" (Desimone et.al., 1984). Gross (1992) concludes that " ...IT cells tend to respond at different rates to a variety of different stimuli." Thus it seems that multidimensional units with Gaussian-like tuning are not only biologically plausible, but ubiquitous in cortical physiology. This claim is not meant to imply that for every feature dimension of a multidimensionally tuned neuron, neurons feeding into it can be found individually tuned to that dimension. For example, for some motion-selective cells in MT the selectivities to spatial frequency and temporal frequency cannot be separated. Yet for these, it may be inappropriate to consider time and space as two independent dimensions and more appropriate to consider velocity as the single dimension in which the neuron is tuned. On the other hand, it is well known that at lower levels in the visual system there do exist cells broadly tuned individually to spatial frequency, orientation, and wavelength, for example, and from these dimensions many complex features can be constructed.

We also observe that not all MBMs have the same applicability in describing properties of cortical neurons. In particular, tuned neurons seem to behave more like Gaussian HBF units than like the sigmoidal units typically found in multilayer perceptrons (MLPs): the tuned response function of cortical neurons resembles $exp(-(||\mathbf{x} - \mathbf{t}||)^2/2\sigma^2$ more than it does $\sigma(\mathbf{x}\dot{\mathbf{w}})$, where σ is a sigmoidal "squashing" function and we define $\mathbf{w}$ as the vector of connection weights including the bias parameter θ. (The typical sigmoidal response to contrast that most neurons display may be treated as a Gaussian of large σ.) For example, when the stimulus to an orientation-selective cortical neuron is changed from its optimal value in any direction, the neuron's response typically decreases. The activity of a Gaussian HBF unit would also decline with any change in the stimulus away from its optimal value $\mathbf{t}$. But for the sigmoid unit certain changes away from the optimal stimulus will not decrease its activity, for example when the input $\mathbf{x}$ is multiplied by a constant $\alpha > 1$.

Lastly, we observe that although the Gaussian is the simplest and most readily interpretable RBF in physiological terms, it might not ultimately provide the best fit to all the physiological data once in. In espousing the general theory of MBMs for cortical mechanisms of object recognition, we do not confine ourselves to Gaussian RBFs as the only model of cortical neurons, but only at present the most plausible.

3.2.2 Centers and a fundamental property of our sensory world

We can recognize almost any object from any of many small subsets of its features, visual and non-visual. We can perform many motor actions in several different ways. In most situations, our sensory and motor worlds are *redundant.* In the language of the previous section this means that instead of high-dimensional centers any of *several lower dimensional centers are often sufficient* to perform a given task. This means that the "and" of a high-dimensional conjunction can be replaced by the "or" of its components – a face may be recognized by its eyebrows alone, or a mug by its colour. To recognize an object, we may use not only templates comprising all its features, but also subtemplates, comprising subsets of features (and in fact exemplary sets of centers capable of generating most eyes, say). This is similar in spirit to the use of several small templates as well as a whole-face template in the Brunelli-Poggio work on frontal face recognition (Brunelli and Poggio, 1992).

Splitting the recognizable world into its additive parts may well be preferable to reconstructing it in its full multidimensionality, because a system composed of several independently accessible parts is inherently more robust than a whole, simultaneously dependent on each of its parts. The small loss in uniqueness of recognition is easily offset by the gain against noise and occlusion. This reduction of the recognizable world into its parts may well be what allows us to "understand" the things that we see (see Appendix B).

3.2.3 How many cells?

The idea of sparse population coding is consistent with much physiological evidence, beginning even at the retinal level where colors are coded by 3 types of photoreceptors. Young and Yamane (1992) conclude from neurophysiological recordings of IT cells broadly tuned to physical prototypes of faces: *"Rather than representing each cell as a vector in the space, the cell could be represented as a surface raised above the feature space. The height of the surface above each point in the feature space would be given by the response magnitude of the cell to the corresponding stimuli and population vectors would be derived by summing the response weighted surfaces for each cell for each stimulus."* MBMs also suggest that the importance of the object and the exposure to it may determine how many centers are devoted to its recognition. Thus faces may have a more "punctate" representation than other objects simply because more centers are used. Psychophysical experiments do suggest that an increasing number of centers is created under extended training to recognize a 3D object (Bülthoff and Edelman, 1992).

While we would not dare to make a specific prediction on the absolute number of cells used to code for a specific object, computational experiments and our arguments here suggest at least a minimum bound. Simulations by Poggio and Edelman (1990) suggest that in an MBM model a minimum of 10-100 units is needed to represent all possible views of a 3D object. We think that the primate visual system could not achieve the same representation with fewer than on the order of 1000. This number seems physiologically plausible, although we expect that the actual number will depend strongly on the reliability of the neurons, training of the animal, relevance of the represented object and other properties of the implementation. Thus we envisage that training a monkey to one view of a target object may "create" at least on the order of 100 cells tuned to that view [1] in the relevant cortical area, with a generalization field similar to the one shown in figure 5. Training to an additional view may *create or recruit* cells tuned to that view. Overtraining a monkey on a specific object should result in an over-representation in cortex of that object – more cells than normally expected would be tuned to views of the object. Recent results from Kobatake, et. al. (1993) suggest that up to two orders of magnitude more cells may be "created" in IT (or, rather, the stimulus selectivities of existing cells altered) on over-training to specific objects.

Note that we do not mean to imply that *only* 10 - 1000 cortical cells would be active on presentation of an object. Many more would be activated than those that are critical for its representation. We suggest only that the activity of approximately 100 cells should be sufficient to discriminate between two distinct objects. This conclusion is broadly supported by the conclusion of Young and Yamane (1992) that the population response of approximately 40 cells in IT is approximately sufficient to encode a particular face, and by the related observation of Britten, et.al. (1992) that the activity of a small pool of weakly correlated neurons in MT is sufficient to predict a monkey's behavioral response in a motion detection task.

3.2.4 HBF centers and biophysical mechanisms

How might multidimensional Gaussian receptive fields be synthesized from known receptive fields and biophysical mechanisms?

The simplest answer is that cells tuned to complex features are constructed from a hierarchy of simpler cells tuned to incrementally larger conjunctions of elementary features. This idea – a standard explanation – can immediately be formalized in terms of Gaussian radial basis functions, since a multidimensional Gaussian function can be decomposed into the product of lower dimensional Gaussians (Marr and Poggio, 1977; Ballard, 1986; Mel, 1988; Poggio and Girosi, 1990b).

The scheme of figure 6 is a possible example of an implementation of Gaussian Radial Basis functions in terms of physiologically plausible mechanisms. The first step applies to situations in which the inputs are place-coded, that is, in which the value of the input is represented by its location in a spatial array of cells – as, for example, the image coordinates x, y are encoded by the spatial pattern of photoreceptor activites. In this case Gaussian radial functions in one, two and possibly three dimensions can be implemented as *receptive fields* by weighted connections from the sensor arrays (or some retinotopic array of units whose activity encodes the location of features). If the inputs are interval-coded, that is, if the input value is represented by the continuously-varying firing rate of a single neuron, then a one-dimensional Gaussian-like tuned cell can be created by passing the input value through multiple sigmoidal functions with different thresholds and taking their difference.

Consider, for example, the problem of encoding colour. At the retinal level, colour is recorded by the triplet of activities of three types of cell: the cone-opponent red-green (R-G) and blue-yellow (B-Y) cells and the luminance (L) cell. An R-G cell signals increasing amounts of red or decreasing amounts of green by increasing its firing rate. Thus it does not behave like a Gaussian tuned cell. But at higher levels in the visual system, there exist cells that behave very much like units tuned to particular values in 3D colour space (Schein and Desimone, 1990). How are these multidimensional tuned colour cells constructed from one-dimensional rate-coded cells? We suggest that one-dimensional Gaussian tuned cells may be created by the above mechanism, selective to restricted ranges of the three colour axes.

Gaussians in higher dimensions can then be synthesized as products of one and two dimensional receptive fields. An important feature of this scheme is that the multidimensional radial functions are synthesized directly by appropriately weighted connections from the sensor arrays, without any need of an explicit computation of the norm and the exponential. From this perspective the computation is performed by *Gaussian receptive fields* and their combination (through some approximation to multiplication), rather than by threshold functions. The view is in the spirit of the key role that the concept of receptive field has always played in neurophsyiology. It *predicts a sparse population coding* in terms of low-dimensional feature-like cells and mul-

[1] Probably in different ways: different cells may be tuned to different parts of the view and may converge to different "prototypes" representing that component; when we use the term "prototype" we have in mind the "caricatures" of Brunelli and Poggio

tidimensional Gaussian-like receptive fields, somewhat similar to template-like cells, a prediction that could be tested experimentally on cortical cells.

The multiplication operation required by the previous interpretation of Gaussian RBFs to perform the "conjunction" of Gaussian receptive fields is not too implausible from a biophysical point of view. It could be performed by several biophysical mechanisms (see Koch and Poggio, 1987; Poggio, 1990). Here we mention several possibilities:

1. inhibition of the silent type and related synaptic and dendritic circuitry (see Poggio and Torre, 1978; Torre and Poggio, 1978).
2. the AND-like mechanism of NMDA receptors
3. a logarithmic transformation, followed by summation, followed by exponentiation. The logarithmic and exponential characteristic could be implemented in appropriate ranges by the sigmoid-like pre-to-postsynaptic voltage transduction of many synapses.
4. approximation of the multiplication by summation and thresholding as suggested by Mel (1990).

If the first or second mechanism is used, the product of figure 6 can be performed directly on the dendritic tree of the neuron representing the corresponding radial function. In the case of Gaussian receptive fields used to synthesize Gaussian radial basis functions, the center vector is effectively stored in the position of the 2D (or 1D) receptive fields and in their connections to the product unit(s). This is plausible physiologically.

Linear terms (direct connections from the inputs to the output) can be realized directly as inputs to an output neuron that summates linearly its synaptic inputs. An output nonlinearity such as a threshold or a sigmoid or a log transformation may be advantageous for many tasks and will not change the basic form of the model (see Poggio and Girosi, 1989).

3.2.5 Circuits

There is at least one other way to implement HBFs networks in terms of known properties of neurons. It exploits the equivalence of HBFs with MLP networks for normalized inputs (Maruyama et. al., 1992). If the inputs are normalized (as usual for unitary input representations), an HBF network could be implemented as a MLP network by using threshold units. There is the problem, though, in normalizing the inputs in a biologically plausible way. MLP networks have a straightforward implementation in terms of linear excitation and inhibition and of the threshold mechanism of the spike for the sigmoidal nonlinearity. The latter could also be implemented in terms of the pre-postsynaptic relationship between presynaptic voltage and postsynaptic voltage. In either case this implementation requires one neuron per sigmoidal unit in the network.

Mel (1992) has simulated a specific biophysical hypothesis about the role of cortical pyramidal cells in implementing a learning scheme that is very similar to a HBF network. Marr (1970) had proposed another similar model of how pyramidal cells in neocortex could learn to discriminate different patterns. Marr's model is, in a sense, the look-up table limit of our HBF model.

3.3 Mechanisms for learning

Reasoning from the HBF model, we expect two mechanisms for learning, probably with different localizations, one that could occur unsupervised and thus is similar to adaptation, and one supervised and probably based on Hebb-like mechanisms.

The first stage of learning would occur at the site of the *centers.* Let us remember that a center represents a neuron tuned to a particular visual stimulus, for example, a vertically oriented light bar. The coefficients c_α represent the synaptic weights on the connections that the neuron makes to the output neuron that registers the network's response. In the simple RBF scheme the only parameters updated by learning are these coefficients. But in constructing the network, the centers must be set to values equal to the input examples. Physiologically, then, selecting the centers $\mathbf{t}_\alpha$ might correspond to choosing or re-tuning a subset of neurons selectively responsive to the range of stimulus attributes encountered in the task. This stage would be very much like *adaptation*, an adjustment to the prevailing stimulus conditions. It could occur *unsupervised*, and would strictly depend only on the stimuli, not on the task. Of course we would expect some centers to be pretuned by evolution, evn in IT cortex.

The second stage, updating of the coefficients c_α, could occur only *supervised*, since it depends on the full input and output example pairs, or, in other words, on the task. It could be achieved by a simple Hebb-type rule, since the gradient descent equations for the c are (Poggio and Girosi, 1989):

$$\dot{c}_\alpha = \omega \sum_{i=1}^{N} \Delta_i G(\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2_{\mathbf{W}}) \quad , \qquad (5)$$

with $\alpha = 1, \ldots, n$ and Δ_i is the squared error between the correct output for example i and the actual output of the network. Thus equation 5 says that the change in the c_α should be proportional to the product of the activity of the unit i and the output error of the network. In other words, the "weights" of the c synapses will change depending on the product of pre- and postsynaptic activity (Poggio and Girosi, 1989; Mel, 1988; Mel, 1990).

In the RBF case, the centers are fixed after they are initially selected to conform to the input examples. In the HBF case, the centers move to optimal locations during learning. This movement may be seen as task-specific or *supervised* fine-tuning of the centers' stimulus selectivities. It is highly unlikely that the biological visual system chooses between distinct RBF-like and HBF-like implementations for given problems. It is possible, though, that tuning of cell selectivities can occur in at least two different ways, corresponding to the *supervised* and *unsupervised* stages outlined here. We might also expect that these two types of learning of "centers" could occur on two different time scales: one fast, corresponding to selecting centers from a pre-existing set, and one slow, corresponding to synthesizing new centers or refining their stimulus specificities. The cortical locations of these two mechanisms, one unsupervised, the other supervised, may be different and have interesting implications on how to interpret data on transfer of learning (see Poggio, Fahle and Edelman, 1992).

For fast, unsupervised learning, there might be a large reservoir of centers already available, most of them with an associated $c = 0$, as suggested by Mel (1990) in a slightly different context. The relevant ones would gain a non-zero weight during the adaptive process. Alternatively, the mechanism could be similar to some of the unsupervised learning models described by Linsker (1990), Intrator and Cooper (1991), Földiak (1991) and others.

Slow, supervised learning may occur by the stabilization of electrically close synapses depending on the degree to which they are co-activated (see, e.g. Mel, 1992). In this scheme, the changes will be formation and stabilization of synapses and synapse clusters (each synapse representing a Gaussian field) on a cortical pyramidal cell simply due to correlations of presynaptic activities. We suggest that this synthesis of new centers, as would be needed in learning to recognize unfamiliar objects, should be slower than selecting centers from an existing pool. But some recent data on perceptual learning (e.g. Fiorentini and Berardi, 1981; Poggio, Fahle and Edelman, 1992; Karni and Sagi, 1990) indicates otherwise: the fact that human observers rapidly learn entirely novel visual patterns suggests that new centers might be synthesized rapidly.

It seems reasonable to conjecture, though, that updating of the elements of the $\mathbf{W}$ matrix may take place on a much slower time scale.

Do the update schemes have a physiologically plausible implementation? Methods like the random-step method (Caprile and Girosi, 1990), that do not require calculation of derivatives, are biologically the most plausible. (In a typical random-step method, network weight changes are generated randomly under the guidance of simple rules; for example, the rule might be to double the size of the random change if the network performace improves and to halve the size if it does not.) In the Gaussian case, with basis functions synthesized through the product of Gaussian receptive fields, moving the centers means establishing or erasing connections to the product unit. A similar argument can be made also about the learning of the matrix $\mathbf{W}$. Notice that in the diagonal Gaussian case the parameters to be changed are exactly the σ of the Gaussians, i.e., the spread of the associated receptive fields. Notice also that the σ for all centers on one particular dimension is the same, suggesting that the learning of w_i may involve the modification of the scale factor in the input arrays rather than a change in the dendritic spread of the postsynaptic neurons. In all these schemes the real problem consists in how to provide the "teacher" input.

4 Predictions and Remarks

To summarize, we highlight the main predictions made by our interpretation of Memory-Based Models of the brain.

Predictions:

1. **Sparse population coding**. The general issue of how the nervous system represents objects and concepts is of course unresolved. "Sparse" or "punctate" coding theories propose that individual cells are highly specific and encode individual patterns. "Population" theories propose that distributed activity in a large number of cells underlies perception. Models of the HBF type suggest that a small number of cells or groups of cells (the centers), each broadly tuned, may be sufficient to represent a $3D$ object. Thus our interpretation of MBMs *predicts* a "sparse population coding", partway between fully distributed representations and grandmother neurones. Specifically, we predict that the activity of approximately 100 cells is sufficient to distinguish any particular object, although many more cells may be active at the same time.

2. **Viewer-centered and object-centered cells**. Our model (see the module of Figure 7) predicts the existence of viewer-centered cells (the centers) and object-centered cells (the output of the network). Evidence pointing in this direction in the case of face cells in IT is already available. We predict a similar situation for other $3D$ objects. It should be noted that the module of Figure 7 is only a small part of an overall architecture. We predict the existence of other types of cells, such as pose-tuned, expression-tuned and illumination-tuned cells. Very recently N. Logothetis (pers. comm.) has succeeded in training monkeys to recognize the same objects used in human psychophysics, and has reproduced the key results of Bülthoff and Edelman (1992). He also succeeded in measuring generalization fields of the type shown in figure 5 after training on a single view. We believe that such a psychophysically measured generalization field corresponds to a group of cells tuned in a Gaussian-like manner to that view. We expect that in trained monkeys, cells exist corresponding to the hidden units of a HBF network, specific for

the training view, as well as possibly other cells responding to subparts of the view. We conjecture (although this is not a critical prediction of the theory) that the step of creating the tuned cells, i.e. the centers, is unsupervised: in other words, that to create the centers it would be sufficient to expose monkeys to target views without actually training them to respond in specific ways.

3. **Cells tuned to full views and cells tuned to parts**. Our model implies that both high-dimensional and low-dimensional centers should exist for recognizable objects, corresponding to full templates and template parts. Physiologically this corresponds to cells that require the whole object to respond (say a face) as well as cells that respond also when only a part of the object is present (say, the mouth).

4. **Rapid Synaptic plasticity**. We predict that the formation of new centers and the change in synaptic weights may happen over short time scales (possibly minutes) and relatively early in the visual pathway (see Poggio, Fahle and Edelman, 1992). As we mentioned, it is likely that the formation of new centers is unsupervised while other synaptic changes, corresponding to the c_i coefficients, should be supervised.

5 HBF-like modules and theories of the brain

As theories of the brain (or of parts of it) HBFs networks replace computation with memory. They are equivalent to modules that work as *interpolating look-up tables*. In a previous paper one of us has discussed how theories of this type can be regarded as a modern version of the "grandmother cell" idea (Poggio, 1990).

The proposal that much information processing in the brain is performed through modules that are similar to *enhanced look-up tables* is attractive for many reasons. It also promises to bring closer apparently orthogonal views, such as the *immediate perception* of Gibson (1979) and the *representational theory* of Marr (1982), since almost iconic "snapshots" of the world may allow the synthesis of computational mechanisms equivalent to vision algorithms. The idea may change significantly the computational perspective on several vision tasks. As a simple example, consider the different specific tasks of hyperacuity employed by psychophysicists. The proposal would suggest that an appropriate module for the task, somewhat similar to a new "routine," may be synthesized by learning in the brain (see Poggio, Fahle and Edelman, 1992).

The claim common to several network theories, such as Multilayer Perceptrons and HBF networks, is that the brain can be explained, at least in part, in terms of approximation modules. In the case of HBF networks these modules can be considered as a powerful extension of look-up tables. MLP networks cannot be interpreted directly as modified look-up tables (they are more similar to an extension of multidimensional Fourier series), but the case of normalized inputs shows that they are similar to using templates.

The HBF theory predicts that population coding (broadly tuned neurons combined linearly) is a consequence of extending a look-up table scheme – corresponding to interval coding – to yield interpolation (or more precisely approximation), that is generalization. In other words, *sparse population coding* and *neurons tuned to specific optimal stimuli* are direct and strong predictions of HBF schemes. It seems that the hidden units of HBF models bear suggestive similarities with the usual descriptions of cortical neurons as being tuned to optimal multidimensional stimuli. It is of course possible that a hierarchy of different networks – for example MLP networks – may lead to tuned cells similar to the hidden units of HBF networks.

Acknowledgements This paper describes research done within the Center for Biological Information Processing in the Department of Brain and Cognitive Sciences, and at the Artificial Intelligence Laboratory. This research is sponsored by a grant from NATO to Dr. Hurlbert and Dr. Poggio; by a grant from the Office of Naval Research, Cognitive and Neural Sciences Division; by the Office of Naval Research contract N00014–89–J–3139 under the DARPA Artificial Neural Network Technology Program and by a grant from the National Science Foundation under contract ASC-9217041 (including funds from DARPA provided under the HPCC program). Support for the A.I. Laboratory's artificial intelligence research is provided by ONR contract N00014–91–J–4038. Tomaso Poggio is supported by the Uncas and Helen Whitaker Chair at the Whitaker College, Massachusetts Institute of Technology. Anya Hurlbert's research is supported by the JCI (ESRC, SERC and MRC), SERC III, and the Royal Society, England.

A An architecture for recognition: the classification and indexing route to recognition

Here we elaborate on the architecture for a recognition system introduced in Section 2. Figure 1 illustrates the main components of the architecture and its two interlocking routes to recognition. The first route, which we call the classification and indexing route, is essentially equivalent to an earlier proposal (Poggio and Edelman, 1990) in which a HBF network receives inputs in the form of feature parameters and classifies inputs as same or different from the target object. This is a streamlined route to recognition which requires that the features extracted in the early stages of image analysis be sufficient to enable matching with samples in the database. Its goal may be primarily basic level recognition, but it is also the route that might suit best the search for and recognition of an expected object. In that case it may be used to identify objects (at the subordinate level) whose class membership is known in advance. It consists of 3 main stages:

1. **Image measurements**

   The first step is to compute a *primal image representation*, which is a set of sparse measurements on the image, based on appropriate smoothed derivatives, corresponding to center-surround and directional receptive fields. It can be argued that the (vector) measurements to be considered should be multiple nonlinear functions of differential operators applied to the image at sparse locations (for a discussion of linear and non-linear measurement or "matching" primitives see Appendix in Nishihara and Poggio, 1984). (Similar procedures may involve using Gaussians of different scales and orientations [e.g. Marr and Poggio, 1977], Koenderink's "jets," [Koenderink and VanDoorn, 1990], Gabor filters, or wavelets. A regularized gradient of the image also works well.) We call this array of measurements an M-array; in general, it is a vector-valued array). For recognition of frontal images of faces an M-array as small as 30×30 has been found sufficient to encode an image of initial size 512×512 (Brunelli and Poggio, 1992).

2. **Feature detection and measurements**

   Key features, encoded by the primal measurements, are then found and localized. These features may be specific for a specific object class – for the expected class, if it is known in advance, or for an alternative class considered as a potential match. This step can be regarded as performing a sort of template matching with several appropriate examples; when a face is the object of the search, templates may include eye pairs of different size, pose, and expression. In the HBF case the templates would effectively correspond to different centers, and matching would proceed in a more sophisticated way than direct comparison. It is clear that this step may by itself accomplish segmentation. These features may be local or global: they may correspond to eye corners or to mean values of the M-array filtered through a large set of filters.

3. **Classification and indexing**

   Parameter values estimated by the preceding stage for the features of interest – e.g. the distance between eyes and mouth – are used in this stage for classification and indexing in a database of known examples. In many cases this may lead by itself to unique recognition, especially when situational information, such as the expectedness of a particular object, restricts the relevant data base. Classification could be done via a number of classical schemes such as Nearest Neighbor or with modules that are more biologically plausible such as HBF networks.

Some open questions remain:

- What are the features used by the human visual system in the feature detection stage? The "non-local" hypothesis is that there is a large set of filters tuned to different 2D shape features and efficiently doing a kind of template matching on the input. Some functional of the correlation function is then evaluated (such as the max of the correlation or some robust statistics on the correlation values, see Viola and Poggio, in preparation). The results may become some of the components (for that particular filter, i.e. template) of the input vector to object-specific networks consisting of hidden units each tuned to a view and an output unit which is view-invariant. Networks of this type may also exist not only for specific objects but also for general object components, perhaps similar to more precise versions of some of Biederman's geons (Biederman, 1987). They would be synthesized by familiarity and their output may have a varying degree of view invariance depending on the type and number of the tuned cells in the hidden layer. Networks of this type, tuned to a particular shape, could easily be combined conjunctively to represent more complex shapes (but still exploiting the fundamental property of additivity). This general "non-local" scheme avoids the correspondence problem since the components of the input vectors are statistics taken over the whole image, rather than individual pixel values or feature locations. It may well be that – in the absence of a serial mechanism such as eye motions and attentional shifts – the visual system does not have a way to keep and use spatial relations between different components or feaures in an image and that it can only detect the likely "presence" of, say, a few hundred features of various complexity.

- The architecture has to be hierarchical, consisting of a hierarchy of HBF-like networks. For instance, an eye-recognizing MBM network may provide some of the inputs to a face recognition network that will combine the presence (and possibly relative position) of eyes with other face features (remember that a MBM network can be regarded as a disjunction of conjunctions). The inputs to the eye-recognizing networks may be themselves provided by other RBF-like networks; this is similar to the use in the eye-recognizing networks of inputs that are the result of filtering the image through of a few basic filters out of a large vocabulary consisting of hundreds of "elementary" templates, representing a vocabulary of shapes of the type described by Fujita and Tanaka (1992). The description of Perrett and Oram (1992) is consistent with this scenario. At various stages in this hierarchy more invariances may be achieved for position, rotation, scaling etc. in a similar way to how complex cells are built from simple ones.

B An architecture for recognition: the visualization route to recognition

The second potential route to recognition takes a necessary detour from the first route to fine-tune the matching mechanisms. Like the classification pathway it begins with the two stages of *image measurement* and *feature detection*, but diverges because it allows for the possibility that a match between the database and measured image features might not directly be found. Further processing may take place on the image or on the stored examples to bring the two into registration or to narrow the range of the latter. The main purpose of this loop is to correct for deformations before comparing image to data base.

Computational arguments (Breuel, 1992) suggest that this route should separate transformations to be applied to the image (to correct image-plane deformations such as image-plane translations, scaling and rotations) from those to be applied to the database model (which may include rotations-in-depth, illumination changes, and alterations in facial expression, for example). The system may try a number of transformations in parallel and on multiple scales of spatial resolution (see van Essen and Anderson, this volume) until it finds the one that succeeds. In general the whole process may be iterated several times before it achieves a satisfactory level of confidence. In the primate visual system, the likely site for the latter transformations is cortical area IT, whereas the former would probably take place earlier, as available results on properties of IT seems to suggest (Gross, 1992; Perrett et.al., 1982; Perrett and Harries, 1988; Perrett et. al., 1989). The main steps of this hypothetical second route to recognition are:

1. **Image measurement**
2. **Feature detection**
3. **Image rectification**

   The feature detection stage provides information about the location of key features that is used in this stage to normalize for image-plane translation, scaling and image-plane rotation of the input M-array.
4. **Pose estimation**

   3-D pose (2 parameters), illumination, and other parameters (such as facial expression) are estimated from the M-array. This computation could be performed by an MBM module that has "learned" the appropriate estimation function from examples of objects of the same class.
5. **Visualization**

   The models (M-arrays in the data-base corresponding to known objects) are warped in the dimensions of pose and expression and illumination, to bring them into register with the estimate obtained from the input image. The transformation of the models is performed by exploiting information specific to the given object (several views per object may have been stored in memory) or by applying a generic transformation (e.g., for a face, from "serious" to "smiling") learned from objects of the same class. Several transformations may be attempted at this stage before a good match is found in the next step.
6. **Verification and indexing**

   The rectified "image" is compared with the warped data base of standard representations. Open questions remain on how the data base may be organized and what are the most efficient means of indexing it.

References

[1] D. H. Ballard. Cortical connections and parallel processing: structure and function. *Behavioral and Brain Sciences*, 9:67–120, 1986.

[2] R. J. Baron. Mechanisms of human facial recognition. *International Journal of Man Machine Studies*, 15:137–178, 1981.

[3] R. Basri and S. Ullman. Recognition by linear combinations of models. Technical report, The Weizmann Institute of Science, 1989.

[4] Martin Bichsel. *Strategies of Robust Object Recognition for the Identification of Human Faces.* PhD thesis, Eidgenossischen Technischen Hochschule, Zürich, 1991.

[5] I. Biederman. Recognition by components: a theory of human image understanding. *Psychol. Review*, 94:115–147, 1987.

[6] T. M. Breuel. *Geometric aspects of visual object recognition.* PhD thesis, Department of Brain and Cognitive Sciences, Massachussetts Institute of Technology, Cambridge, MA, 1992.

[7] E. Bricolo and H. B. Bülthoff. Translation-invariant features for object recognition. *Perception*, 21-S2:59, 1992.

[8] K. H. Britten, M. N. Shadlen, W. T. Newsome, and J.A. Movshon. The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12:4745–4765, 1992.

[9] R. Brunelli and T. Poggio. HyperBF networks for gender classification. In *Proceedings Image Understanding Workshop*, San Mateo, CA, 1992. Morgan Kaufmann Publishers, Inc.

[10] H. H. Bülthoff and S. Edelman. Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, 89:60–64, 1992.

[11] H. H. Bülthoff, S. Edelman, and E. Sklar. Mapping the generalization space in object recognition. *Invest. Ophthalm. Vis. Science Suppl.*, 32(3):996, 1991.

[12] B. Caprile and F. Girosi. A nondeterministic minimization algorithm. A.I. Memo 1254, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, September 1990.

[13] A. Damasio, D. Tranel, and H. Damasio. Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience*, 13:89–109, 1990.

[14] R. DeSimone, T.D. Albright, C.G. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.*, 4:2051–2062, 1984.

[15] S. Edelman. Features of recognition. In *Proc. Intl. Workshop on Visual Form, Capri, Italy*, New York, 1991. Plenum Press.

[16] S. Edelman and H. H. Bülthoff. Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 32:2385–2400, 1992.

[17] A. Fiorentini and N. Berardi. Learning in grating waveform discrimination: specificity for orientation and spatial frequency. *Vision Research*, 21:1149–1158, 1981.

[18] P. Földiak. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.

[19] I. Fujita and K. Tanaka. Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360:343–346, 1992.

[20] J. J. Gibson. *The ecological approach to visual perception.* Houghton Mifflin, Boston, MA, 1979.

[21] M. A. Goodale, A. D. Milner, L. S. Jakobson, and D. P. Carey. A neurological dissociation between perceiving objects and grasping them. *Nature*, 349:154–156, 1991.

[22] C. G. Gross. Representation of visual stimuli in inferior temporal cortex. *Phil. Trans. Royal Soc. B, London*, 1992.

[23] M.E. Hasselmo, E.T. Rolls, G. C. Baylis, and V. Nalwa. Object-centered encoding by by face-selective neurons in the cortex of in the Superior Temporal Sulcus of the monkey. *Experimental Brain Research*, 75:417–429, 1989.

[24] P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[25] A. Hurlbert and T. Poggio. Do computers need attention? *Nature*, 321(12), 1986.

[26] N. Intrator and L. N. Cooper. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17, 1992.

[27] A. Karni and D. Sagi. Human texture discrimination learning — evidence for low-level neuronal plasticity in adults. *Perception*, 19:335, 1990.

[28] E. Kobatake, K. Tanaka, and Y. Tamori. Learning new shapes changes the stimulus selectivity of cells in the inferotemporal cortex of the adult monkey. *Supplement to Investigative Ophthalmology and Visual Science*, 34:814, 1993.

[29] C. Koch and T. Poggio. Biophysics of computational systems: Neurons, synapses, and membranes. In G. M. Edelman, W. E. Gall, and W. M. Cowan, editors, *Synaptic Function*, pages 637–697. Wiley, New York, NY, 1987.

[30] J. J. Koenderink and A. J. van Doorn. Receptive field families. *Biological Cybernetics*, 63:291–297, 1990.

[31] R. Linsker. Perceptual neural organization: some approaches based on network models and information theory. *Ann. Rev. Neurosci.*, 13:257–281, 1990.

[32] N.K. Logothetis and E. R. Charles. V4 responses to gratings defined by random dot motion. *Supplement to Investigative Ophthalmology and Visual Science*, 31:90, 1990.

[33] D. Marr. A theory for cerebral neocortex. *Proceedings of the Royal Society of London B*, 176:161–234, 1970.

[34] D. Marr. *Vision*. W. H. Freeman, San Francisco, CA, 1982.

[35] D. Marr and T. Poggio. From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull.*, 15:470–488, 1977.

[36] M. Maruyama, F. Girosi, and T. Poggio. A connection between HBF and MLP. A.I. Memo No. 1291, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[37] B. W. Mel. MURPHY: a robot that learns by doing. In D. Z. Anderson, editor, *Neural Information Processing Systems*. American Institute of Physics, University of Colorado, Denver, 1988.

[38] B. W. Mel. The Sigma-Pi column: A model of associative learning in cerebral neocortex. Computational and Neural Systems Program Memo 6, California Institute of Technology, 1990.

[39] B. W. Mel. NMDA-based pattern-discrimination in a modeled cortical neuron. *Neural Computation*, 4:502–517, 1992.

[40] T. Nazir and J. K. O'Regan. Some results on translation invariance in the human visual system. *Spatial vision*, 5:81–100, 1990.

[41] H.K. Nishihara and T. Poggio. Stereo vision for robotics. In M. Brady, editor, *Robotics Research: the First International Symposium*. The MIT Press, 1984.

[42] D. I. Perrett and M. H. Harries. Characteristic views and the visual inspection of simple faceted and smooth objects: tetrahedra and potatoes. *Perception*, 17:703–720, 1988.

[43] D. I. Perrett, A. J. Mistlin, and A. J. Chitty. Visual neurones responsive to faces. *Trends in Neurosciences*, 10:358–364, 1989.

[44] D. I. Perrett and S. Oram. The neurophysiology of shape processing. *Image and Visual Computing*, 1992. submitted.

[45] D. I. Perrett, E. T. Rolls, and W. Caan. Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.*, 47:329–342, 1982.

[46] D. I. Perrett, P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, and M.A. Jeeves. Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proc. R. Soc. London B*, 223:293–317, 1985.

[47] T. Poggio. A theory of how the brain might work. In *Cold Spring Harbor Symposia on Quantitative Biology*, pages 899–910. Cold Spring Harbor Laboratory Press, 1990.

[48] T. Poggio. 3d object recognition and prototypes: One 2d view may be sufficient. Technical Report 9107-02, I.R.S.T, 1991.

[49] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.

[50] T. Poggio, M. Fahle, and S. Edelman. Fast Perceptual Learning in Visual Hyperacuity. *Science*, 256:1018–1021, May 1992.

[51] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[52] T. Poggio and F. Girosi. Extension of a theory of networks for approximation and learning: dimensionality reduction and clustering. A.I. Memo 1167, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990a.

[53] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), September 1990b.

[54] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990c.

[55] T. Poggio and V. Torre. A theory of synaptic interactions. In W.E. Reichardt and T. Poggio, editors, *Theoretical approaches in neurobiology*, pages 28–38. The M.I.T Press, Cambridge, MA, 1978.

[56] T. Poggio and T. Vetter. Recognition and structure from one 2D model view: observations on prototypes, object classes and symmetries. A.I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[57] D. G. Purcell and A.L. Stewart. The face-detection effect: Configuration enhances detection. *Perception and Psychophysics*, 43:355–366, 1988.

[58] S. J. Schein and R. Desimone. Spectral properties of V4 neurons in the macaque. *Journal of Neuroscience*, 10:3369–3389, 1990.

[59] Luigi Stringa. Eyes Detection for Face Recognition. Technical Report 9203-07, I.R.S.T, 1992a.

[60] Luigi Stringa. Automatic Face Recognition using Directional Derivatives. Technical Report 9205-04, I.R.S.T, 1992b.

[61] M. J. Swain and D. H. Ballard. Indexing via color histograms. In *Proceedings of the International Conference on Computer Vision*, pages 390–393, Osaka, Japan, 1990. IEEE.

[62] V. Torre and T. Poggio. A synaptic mechanism possibly underlying directional selectivity to motion. *Proc. R. Soc. Lond. B*, 202:409–416, 1978.

[63] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[64] T. Vetter, T. Poggio, and H. B. Bülthoff. 3d object recognition: Symmetry and virtual views. Artificial Intelligence Laboratory Memo 1409, Massachusetts Institute of Technology, 1992.

[65] M.P. Young and S. Yamane. Sparse population coding of faces in the inferotemporal cortex. *Science*, 256:1327–1331, 1992.

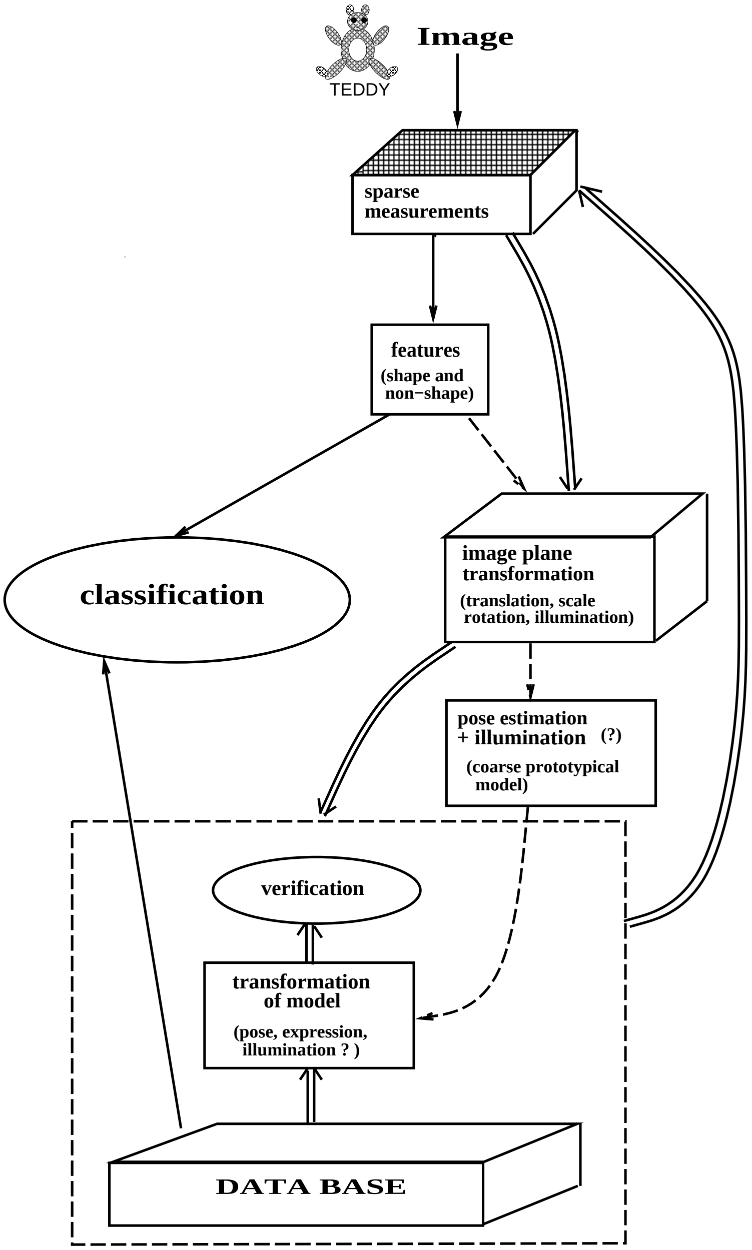
TEDDY
Image
sparse measurements
features (shape and non−shape)
classification
image plane transformation (translation, scale rotation, illumination)
pose estimation + illumination (?) (coarse prototypical model)
verification
transformation of model (pose, expression, illumination ?)
DATA BASE

Figure 1: *A sketch of an architecture for recognition with two hypothetical routes to recognition. Single arrows represent the classification and indexing route described in Appendix A. Double arrows represent the main visualization route, and dashed arrows alternative pathways within it.*

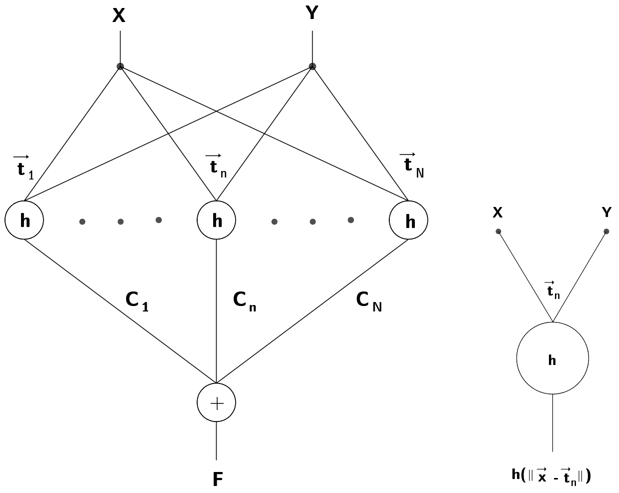

Figure 2: *A RBF network for the approximation of two-dimensional functions (left) and its basic "hidden" unit (right).* $\mathbf{x}$ *and* $\mathbf{y}$ *are components of the input vector which is compared via the RBF* $\mathbf{h}$ *at each center* $\mathbf{t}$*. Outputs of the RBFs are weighted by the* $\mathbf{c_i}$ *and summed to yield the function* $\mathbf{F}$ *evaluated at the input vector.* N *is the total number of centers.*

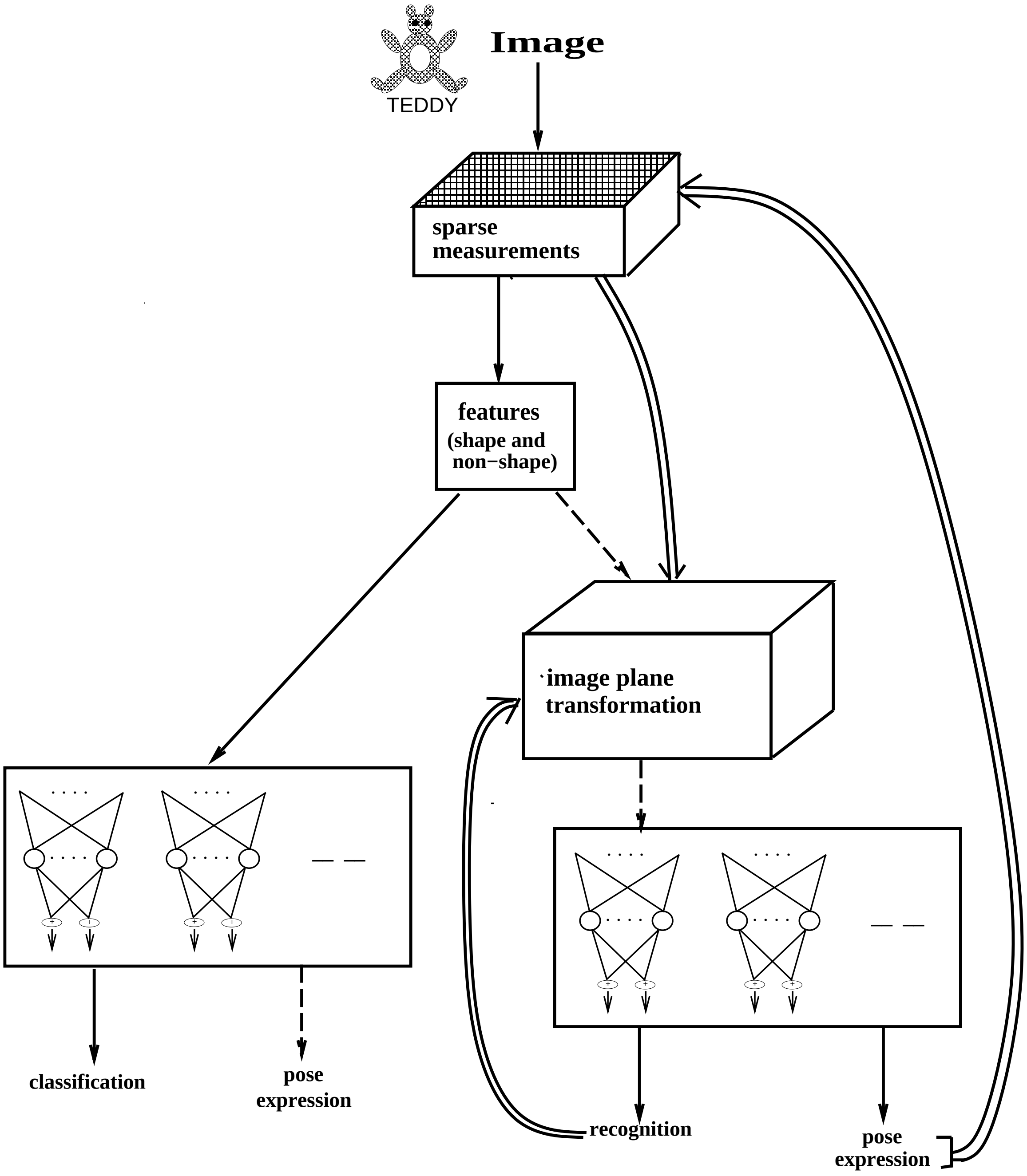

Figure 3: *A sketch of possibly the most compact (but not the only!) implementation of the proposed recognition architecture in terms of modules of the HBF type.*

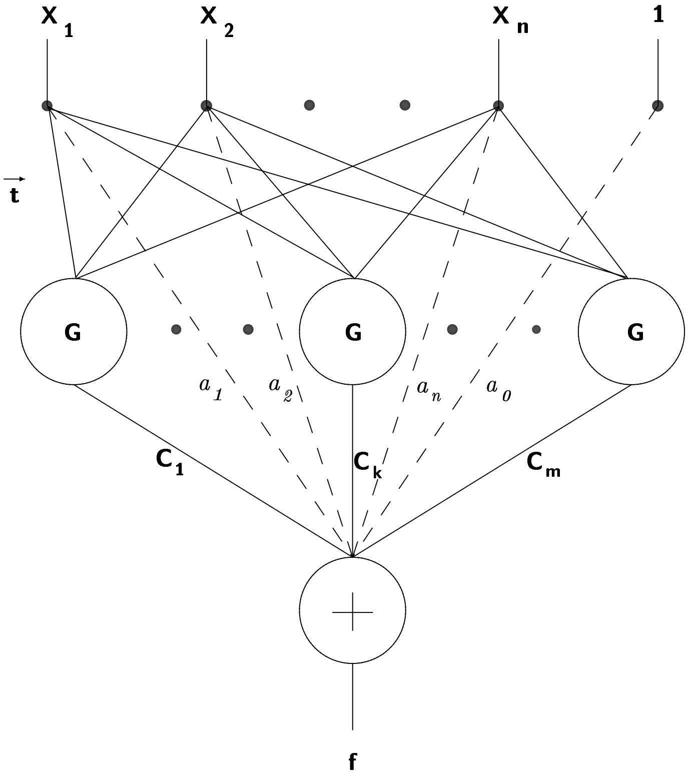

Figure 4: *A network of the Hyper Basis Functions type. For object recognition the inputs could be image measurements such as values of different filters at each of a number of locations in the image. The network is a natural extension of the template matching scheme and contains it as a special case. The dotted lines correspond to linear and constant terms in the expansion. The output unit may contain a sigmoidal transformation of the sum of its inputs (see Poggio and Girosi, 1990b).*

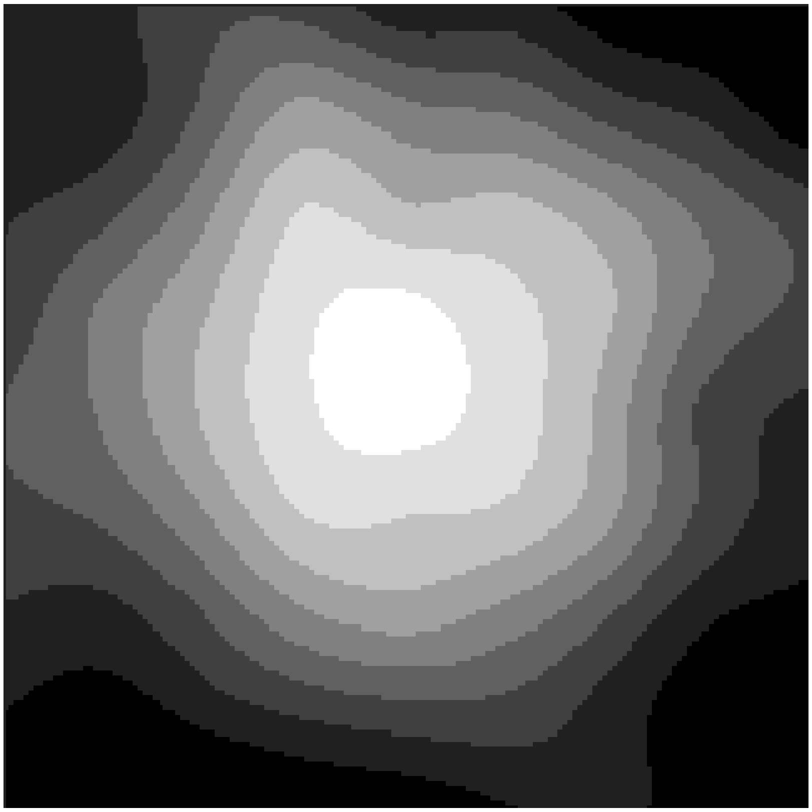

Figure 5: *The generalization field associated with a single training view. Whereas it is easy to distinguish between, say, tubular and amoeba-like* 3*D objects, irrespective of their orientation, the recognition error rate for specific objects within each of those two categories increases sharply with misorientation relative to the familiar view. This figure shows that the error rate for amoeba-like objects, previously seen from a single attitude, is viewpoint-dependent. Means of error rates of six subjects and six different objects are plotted vs. rotation in depth around two orthogonal axes (Bülthoff, Edelman and Sklar, 1991; Edelman and Bülthoff, 1992). The extent of rotation was* ±60° *in each direction; the center of the plot corresponds to the training attitude. Shades of gray encode recognition rates, at increments of* 5% *(white is better than* 90%*; black is* 50%*). From Bülthoff and Edelman (1992a). Note that viewpoint independence can be achieved by familiarizing the subject with a sufficient number of training views of the 3D object.*

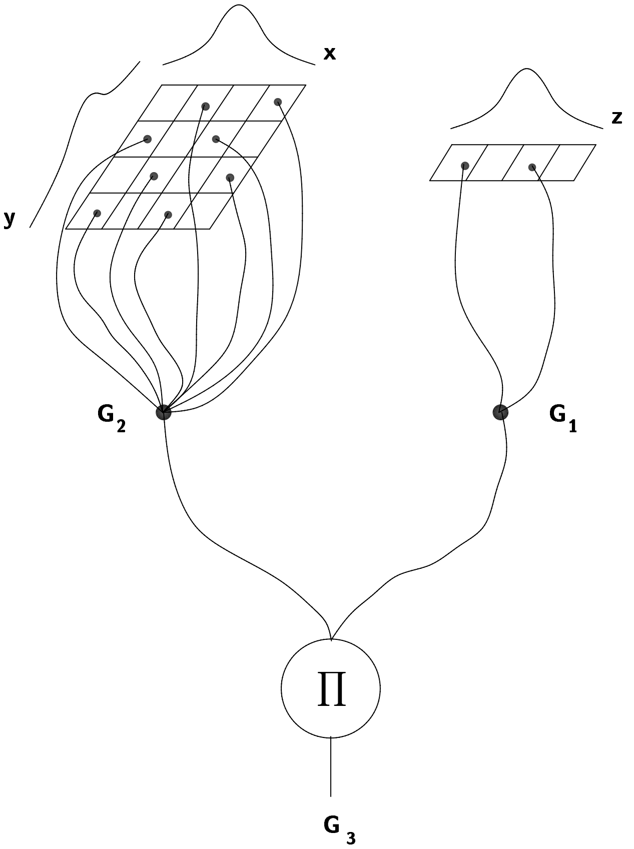

Figure 6: *A three-dimensional radial Gaussian implemented by multiplying a two-dimensional and a one-dimensional Gaussian receptive field . The latter two functions are synthesized directly by appropriately weighted connections from the sensor arrays, as neural receptive fields are usually thought to arise. Notice that they transduce the implicit position of stimuli in the sensor array into a number (the activity of the unit). They thus serve the dual purpose of providing the required "number" representation from the activity of the sensor array and of computing a Gaussian function. 2D Gaussians acting on a retinotopic map can be regarded as representing 2D "features", while the radial basis function represents the "template" resulting from the conjunction of those lower-dimensional features. From Poggio and Girosi (1989a).*

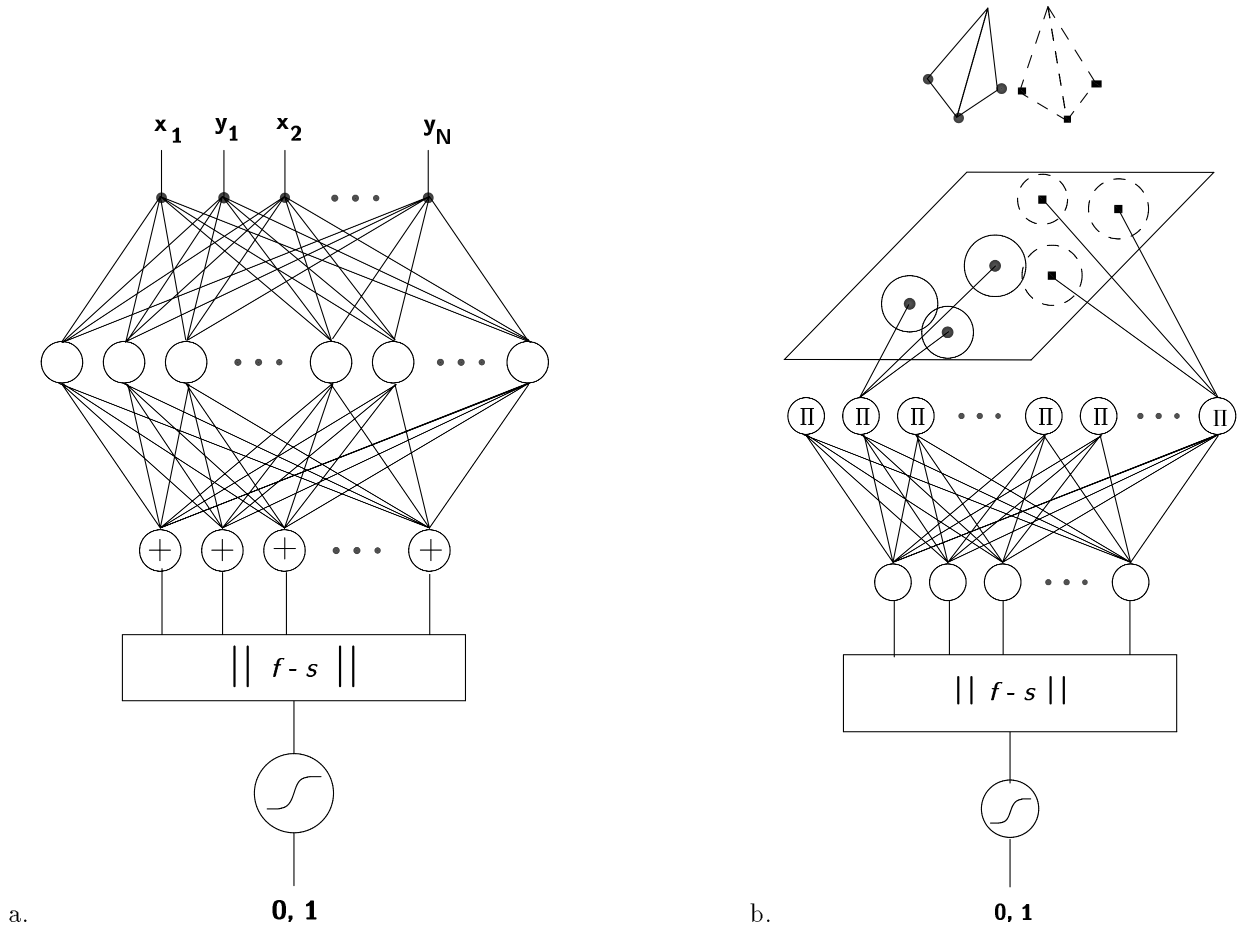

Figure 7: *(a) The HBF network proposed for the recognition of a 3D object from any of its perspective views (Poggio and Edelman, 1990). The network attempts to map any view (as defined in the text) into a standard view, arbitrarily chosen. The norm of the difference between the output vector* $\mathbf{f}$ *and the standard view* $\mathbf{s}$ *is thresholded to yield a* $0, 1$ *answer (instead of the standard view the output of the netwok can be directly a binary classification label). The* $2N$ *inputs accommodate the input vector* $\mathbf{v}$ *representing an arbitrary view. Each of the* n *radial basis functions is initially centered on one of a subset of the* M *views used to synthesize the system* $(n \leq M)$*. During training each of the* M *inputs in the training set is associated with the desired output, i.e., the standard view* $\mathbf{s}$*. Fig. (b) shows a completely equivalent interpretation of (a) for the special case of Gaussian radial basis functions. Gaussian functions can be synthesized by multiplying the outputs of two-dimensional Gaussian receptive fields, that "look" at the retinotopic map of the object point features. The solid circles in the image plane represent the 2D Gaussians associated with the first radial basis function, which represents the first view of the object. The dotted circles represent the 2D receptive fields that synthesize the Gaussian radial function associated with another view. The 2D Gaussian receptive fields transduce values of features, represented implicitly as activity in a retinotopic array, and their product "computes" the radial function without the need of calculating norms and exponentials explicitly. From Poggio and Girosi (1990c).*