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

A.I. Memo No. 1472 April 1994
C.B.C.L. Paper No. 94

View-based Models of 3D Object Recognition and Class-specific Invariances

Nikos K. Logothetis, Thomas Vetter, Anya Hurlbert and Tomaso Poggio

Abstract

This paper describes the main features of a view-based model of object recognition. The model tries to capture general properties to be expected in a biological architecture for object recognition. The basic module is a regularization network in which each of the hidden units is broadly tuned to a specific view of the object to be recognized. The network output, which may be largely view independent, is first described in terms of some simple simulations. The following refinements and details of the basic module are then discussed: (1) some of the units may represent only components of views of the object – the optimal stimulus for the unit, its "center", is effectively a complex feature; (2) the units' properties are consistent with the usual description of cortical neurons as tuned to multidimensional optimal stimuli; (3) in learning to recognize new objects, preexisting centers may be used and modified, but also new centers may be created incrementally so as to provide maximal invariance; (4) modules are part of a hierarchical structure: the output of a network may be used as one of the inputs to another, in this way synthesizing increasingly complex features and templates; (5) in several recognition tasks, in particular at the basic level, a single center using view-invariant features may be sufficient.

Modules of this type can deal with recognition of specific objects, for instance a specific face under various transformations such as those due to viewpoint and illumination, provided that a sufficient number of example views of the specific object are available. An architecture for 3D object recognition, however, must cope – to some extent – even when only a single model view is given. The main contribution of this paper is an outline of a recognition architecture that deals with objects of a *nice* class undergoing a broad spectrum of transformations – due to illumination, pose, expression and so on – by exploiting prototypical examples. A *nice* class of objects is a set of objects with sufficiently similar transformation properties under specific transformations, such as viewpoint transformations. For *nice* object classes, we discuss two possibilities: (a) class-specific transformations are to be applied to a single model image to generate additional *virtual* example views, thus allowing some degree of generalization beyond what a single model view could otherwise provide; (b) class specific, view-invariant features are learned from examples of the class and used with the novel model image, without an explicit generation of virtual examples.

This report describes research done at the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at MIT. This research is sponsored by grants from the Office of Naval Research under contracts N00014-91-J-1270, N00014-92-J-1879 and N00014-93-1-0385; by a grant from the National Science Foundation under contract ASC-9217041 (funds provided by this award include funds from DARPA provided under the HPCC program). Additional support is provided by the North Atlantic Treaty Organization, ATR Audio and Visual Perception Research Laboratories, and Siemens AG. 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 MIT's Whitaker College.

1 Introduction

In the past three years we have been developing systems for 3D object recognition that we label *view-based* (or memory-based, see Poggio and Hurlbert, 1993) since they require units tuned to views of specific objects or object classes.[1] Our work has led to artificial systems for solving toy problems such as the recognition of paperclips as in Figure 3 (Poggio and Edelman, 1990; Brunelli and Poggio, 1991), as well as more real problems such as the recognition of frontal faces (Brunelli and Poggio, 1993; Gilbert and Yang, 1993) and the recognition of faces in arbitrary pose (Beymer, 1993). We have discussed how this approach may capture key aspects of the cortical architecture for 3D object recognition (Poggio, 1990; Poggio and Hurlbert, 1993), we have tested successfully with psychophysical experiments some of the predictions of the model (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992; Schyns and Bülthoff, 1993) and recently we have gathered preliminary evidence that this class of models is consistent with both psychophysics and physiology (specifically, of inferotemporal [IT] cortex) in alert monkeys trained to recognize specific 3D paperclips (Logothetis et al., 1994).

This paper is a short summary of some of our theoretical work; it describes work in progress and it refers to other papers that treat in more detail several aspects of this class of models. Some of these ideas are similar to Perrett's (1989), though they were developed independently from his data; they originate instead from applying regularization networks to the problem of visual recognition and noticing an intriguing similarity between the hidden units of the model and the tuning properties of cortical cells. The main problem this paper addresses is that of how a visual system can learn to recognize an object after exposure to only a *single* view, when the object may newly appear in many different views corresponding to a broad spectrum of image transformations. Our main novel contribution is the outline of an architecture capable of achieving invariant recognition for a single model view, by exploiting transformations learned from a set of prototype objects of the same class.

We will first describe the basic view-based module and illustrate it with a simple simulation. We will then discuss a few of the refinements that are necessary to make it biologically plausible. The next section will sketch a recognition architecture for achieving invariant recognition. In particular, we will describe how it may cope with the problem of recognizing a specific object of a certain class from a single model view. Finally, we will describe an hypothetical, secondary route to recognition – a *visualization* route – in which a) class-specific RBF-like modules estimate parameters of the input image, such as illumination, pose and expression; b) other modules provide the appropriate transformation from prototypes and synthesize a "normalized" view from the input view; c) the normalized input view is compared with the model view in memory. Thus analysis and synthesis networks may be used to close the loop in the recognition process by generating the "neural" imagery corresponding to a certain interpretation and eventually comparing it to the input image. In the last section we will outline some of the critical predictions of this class of biological models and discuss some of the existing data.

[1] Of course the distinction between view-based and object-centered models makes little sense from an information processing perspective: a very small number of views contains full information about the visible 3D structure of an object (compare Poggio and Edelman, 1990). Our view-based label refers to an overall approach that does not rely on an explicit representation of 3D structure and in particular to a biologically plausible implementation in terms of view-centered units.

2 The basic recognition module

Figure 1 shows our basic module for object recognition. As Poggio and Hurlbert (1993) have argued, it is representative of a broad class of memory based modules (MBMs). Classification or identification of a visual stimulus is accomplished by a network of units. Each unit is broadly tuned to a particular view of the object. We refer to this optimal view as the center of the unit. One can think of it as a template to which the input is compared. The unit 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 represents the output of the network.

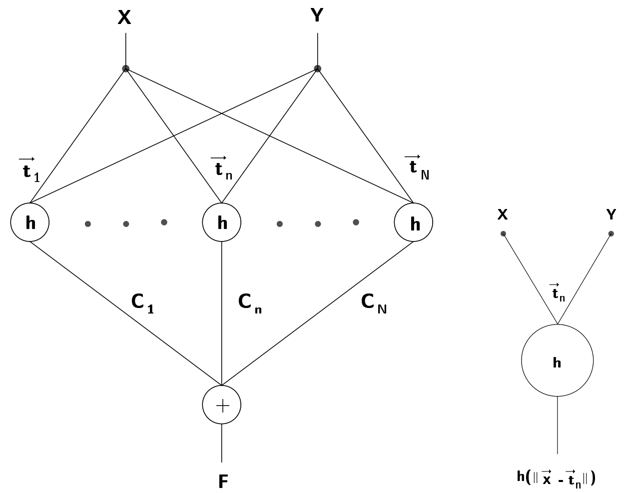

Figure 1: *A RBF network for the approximation of two-dimensional functions (left) and its basic "hidden" unit (right).* $\mathbf{x}$ *and* $\mathbf{y}$ *are components of the input vector which is compared via the RBF* $\mathbf{h}$ *at each center* $\mathbf{t}$*. Outputs of the RBFs are weighted by the* $\mathbf{c_i}$ *and summed to yield the function* $\mathbf{F}$ *evaluated at the input vector. N is the total number of centers.*

Here we consider as an example of such a structure a RBF network that we originally used as a learning network (Poggio and Girosi, 1989) for object recognition while discovering that it was biologically appealing (Poggio and Girosi, 1989; Poggio, 1990; Poggio and Edelman, 1990; Poggio and Hurlbert, 1993) and representative of a much broader class of network architectures (Girosi, Jones and Poggio, 1993).

2.1 RBF networks

Let us review briefly RBF networks. RBF networks are approximation schemes that can be written as (see Figure 1; Poggio and Girosi, 1990b and Poggio, 1990)

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

The Gaussian case, $h(||\mathbf{x} - \mathbf{t}||) = exp(-(||\mathbf{x} - \mathbf{t}||)^2/2\sigma^2)$, is especially interesting:

- 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 $h(||\mathbf{x} - \mathbf{t}||)$.
- In the limiting case of h being a very narrow Gaussian, the network becomes a *look-up* table.
- Centers are like *templates*.

The simplest recognition scheme we consider is the network suggested by Poggio and Edelman (1990) to solve the specific problem of recognizing a particular 3D object from novel views. This is a problem at the *subordinate* level of recognition; it assumes that the object has already been classified on the *basic* level but must be discriminated from other members of its class. In the RBF version of the network, each center stores a sample view of 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.

Consider how the network "learns" to recognize views of the object shown in Figure 3. In this example the inputs of the network are the x, y positions of the vertices of the object images and four training views are used. After training, the network consists of four units, each one tuned to one of the four views as in Figure 2. The weights of the output connections are determined by minimizing misclassification errors on the four views and using as negative examples views of other similar objects ("distractors").

The figure shows the tuning of the four units for images of the "correct" object. The tuning is broad and centered on the training view. Somewhat surprisingly, the tuning is also very selective: the dotted line shows the average response of each unit to 300 similar distractors (paperclips generated by the same mechanisms as the target; for further details about the generation of paperclips see Edelman and Bülthoff, 1992). Even the maximum response to the best distractor is in this case always less than the response to the optimal view. The output of the network, being a linear combination of the activities of the four units, is essentially view-invariant and still very selective. Notice that each center is the *conjunction* of all the features represented: the Gaussian can in fact be decomposed into the product of one-dimensional Gaussians, one for each input component.

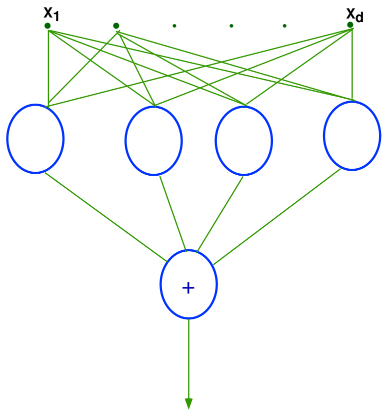

Figure 2: *A RBF network with four units each tuned to one of the four training views shown in the next figure. The tuning curve of each unit is also shown in the next figure. The units are view-dependent but selective relative to distractors of the same type.*

The activity of the unit measures the global similarity of the input vector to the center: for optimal tuning all features must be close to the optimum value. Even the mismatch of a single component of the template may set to zero the activity of the unit. Thus the rough rule implemented by a view-tuned unit is the conjunction of a set of predicates, one for each input feature, measuring the match with the template. On the other hand the output of the network is performing an operation more similar (but not identical because of the eventual output nonlinearity) to the "OR" of the output of the units. Even if the output unit may have a sigmoidal nonlinearity (see Poggio and Girosi, 1990) its output does not need to be zero when one or more of the hidden units are inactive, provided there is sufficient activity in the remaining ones.

This example is clearly a caricature of a view-based recognition module but it helps to illustrate the main points of the argument. Despite its gross oversimplification, it manages to capture some of the basic psychophysical and physiological findings, in particular the existence of view-tuned and view-invariant units and the shape of psychophysically measured recognition fields. In the next section we will list a number of ways in which the network can be made more plausible.

3 Towards more biological recognition modules

The simple model proposed in the previous section contains view-centered hidden units.[2] More plausible versions allow for the centers and corresponding hidden units to be view-invariant if the task requires. In a bio-

[2] A computational reason for why a few views are sufficient can be found in the results (for a specific type of features) of Ullman and Basri (1990). Shashua (1991, 1992) describes an elegant extension of these results to achieve illumination as well as viewpoint invariance.

logical implementation of the network, we in fact expect to find a full spectrum of hidden unit properties, from view-centered to view-invariant. View-centered units are more likely in the case of subordinate level recognition of unfamiliar *not nice* objects (for the definition of a *nice* class, see later); view-invariant units would appear for the basic level recognition of familiar objects. We will now make a number of related observations, some of which can be found in Poggio and Hurlbert (1993), which point to necessary refinements of the model if it is to be biologically plausible.

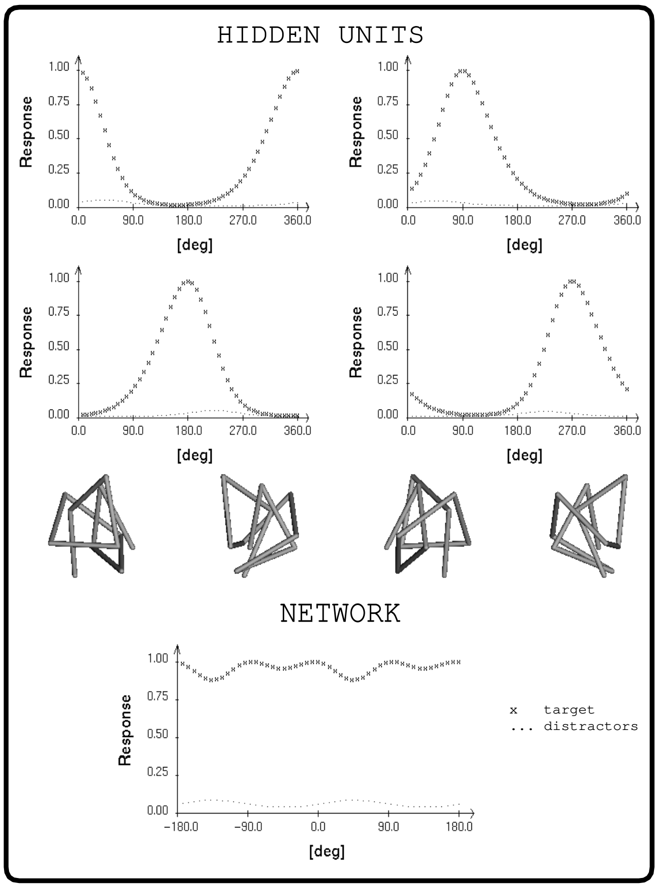

Figure 3: *Tuning of each of the four hidden units of the network of the previous figure for images of the "correct" 3D objects. The tuning is broad and selective: the dotted lines indicate the average response to* 300 *distractor objects of the same type. The bottom graphs show the tuning of the output of the network after learning (that is computation of the weights c): it is view-invariant and object specific. Again the dotted curve indicates the average response of the network to the same* 300 *distractors.*

1. In the previous example each unit has a center which is effectively a full training view. It is much more reasonable to assume that most units in a recognition network should be tuned to *components* of the image, that is to conjunctions of some of the elementary features but not *all* of them. This should allow for sufficient selectivity (the above network performs better than humans) and provide for significant robustness to occlusions and noise (see Poggio and Hurlbert, 1993). 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. Notice that the disjunction (corresponding to the weighted combination of the hidden units) of conjunctions of a small number of features may be sufficient (each conjunction is implemented by a Gaussian center which can be written as the product of one-dimensional Gaussians). To recognize an object, we may use not only templates (i.e. centers in RBF terminology) comprising all its features, but also, and in some cases solely, subtemplates, comprising subsets of features (which themselves constitute "complex" features). This is similar in spirit to the technique of supplementing whole-face templates with several smaller templates in the Brunelli-Poggio work on frontal face recognition (see also Beymer, 1993).

2. The units tuned to complex features mentioned above are similar to IT cells described by Fujita and Tanaka (1992) and could be constructed in a hierarchical way from the output of simpler RBF-like networks. They may avoid the correspondence problem, provided that the system has built-in invariance to image-plane transformations, such as translation, rotation and scaling. Thus cells tuned to complex features are constructed from a hierarchy of simpler cells tuned to incrementally larger conjunctions of elementary features. This idea – popular among physiologists (see Tanaka, 1993; Perrett and Oram, 1993) – 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, 1976; Ballard, 1986; Mel, 1992; Poggio and Girosi, 1990).

3. The features used in the example of Figure 3 (x,y–coordinates of paperclip vertices) are biologically implausible. We have also used other more natural

features such as orientation of lines. An attractive feature of this module is its recursive nature: detection and localization of a line of a certain orientation, say, can be thought of as being performed by a similar network with centers being units tuned to different examples of the desired line type. An eye detector can localize an eye by storing in its units templates of several eyes and using as inputs more elementary features such as lines and blobs. A face recognition network may use units tuned to specific templates of eyes and nose and so on. A homogeneous, recursive approach of this type in which not only object recognition is view-based but also feature localization is view-based has been successfully used in the Beymer-Poggio face recognizer (see Beymer, 1993). Both feature detection and face recognition depend on the use of several templates, the "examples".

4. In this perspective there are probably elementary features such as blobs and oriented lines and center-surround patterns, but there is then a continuum of increasingly complex features corresponding to centers that are conjunctions of more elementary ones. In this sense a center is simply a more complex feature than its inputs and may in turn be the input to another network with even more complex center-features.

5. The RBF network described in the previous sections is the simplest version of a more general scheme (Hyperbasis Functions) given by

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

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*

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

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). 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. 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. As a consequence, Hbf networks may "find" view-invariant features when they exist (Bricolo, in preparation). There are close connections between Hyperbasis Function networks, Multilayer Perceptrons and regularization (see Girosi, Jones and Poggio, 1993).

6. It is also plausible that some of the center-features are "innate", having being synthesized by evolution or by early experience of the individual or more likely by both. We assume that the adult system has at its disposal a vocabulary of simple as well as increasingly more complex center-features. Other centers are synthesized on demand in a task-dependent way. This may happen in the following way. Assume that a network such as the one in Figure 2 has to learn to recognize a new object. It may attempt to do so by using some of the outputs in the pool of existing networks as its inputs. At first no new centers are allocated and only the linear part of the network is used, corresponding to the term $p(\mathbf{x})$ in equation 1 and to direct connections between inputs and output (not shown in Figure 2). This of course is similar to a simple OR of the input features. Learning may be successful in which case only some of the inputs will have a nonzero weight. If learning is not successful – or sufficiently weak – a new center of minimal dimension may be allocated to mimic a component of one of the training views. New centers of increasing dimensionality – comprising subsets of components, up to the full view – are added while old centers are continually pruned until the performance is satisfactory. Centers of dimension 2 effectively detect conjunctions of pairs of input features (see also Mel, 1992). It is not difficult to imagine learning strategies of this type that would select automatically centers, i.e. complex features, that are as view invariant as possible (this can be achieved by modifying the associated parameters c and/or w in the $\mathbf{W}$ matrix). Such features may be global – such as color – but we expect that they will be mostly local and perhaps underlie recognition of geon-like components (see Edelman, 1991 and Biederman, 1987). View-invariant features may be used in basic-level more than in subordinate-levels recognition tasks.

7. One essential aspect of the simplest (RBF) version of the model is that it contains key units which are viewer-centered, not object-centered. This aspect is independent of whether the model is 2D or 3D, a dichotomy which is not relevant here. Each center may consist of a set of features that may mix 2D with 3D information, by including shading, occlusion or binocular disparity information, for example. The features that depend on the image geometry will necessarily be viewpoint-dependent, but features such as color may be viewpoint-independent. As we mentioned earlier, in situations in which view-invariant features exist (for basic as well as for subordinate level recognition) centers may actually be view-independent.

8. The network described here is used as a classifier that performs *identification*, or subordinate-level recognition: matching the face to a stored mem-

ory, and thereby labeling it. A similar network with a different set of centers could perform also basic-level recognition: distinguishing objects that are faces from those that are not.

4 Virtual Views and Invariance to Image Transformations: towards a Recognition Architecture

In the example given above, the network learns to recognize a particular 3D object from novel views and thereby achieves one crucial aim in object recognition: viewpoint invariance. But recognition does not involve solely or simply the problem of recognizing objects in hitherto unseen poses. Hence, as Poggio and Hurlbert (1993) emphasize, the cortical architecture for recognition cannot consist simply of a collection of the modules of Figures 3 and 1, one for each recognizable object. The architecture must be more complex than that cartoon, because recognition must be achieved over a variety of image transformations, not just those due to changes in viewpoint, but also those due to translation, rotation and scaling of the object in the image plane, as well as non-image-plane transformations, such as those due to varying illumination. In addition, the cortex must also recognize objects at the basic as well as subordinate level.

In the network described above, viewpoint invariance is achieved by exploiting several sample views of the specific object. This strategy might work to obtain invariance under other types of transformations also, provided sufficient examples of the object under sample transformations are available. But suppose that example views are not available. Suppose that the visual system must learn to recognize a given object under varying illumination or viewpoint, starting with only a single sample view. This is the problem that we will focus on in the next few sections, that of subordinate level recognition under non-image-plane transformations, given only a single model view.

Probably the most natural solution is for the system to exploit certain invariant features, learned from examples of objects of the same class. These features could supplement the information contained in the single model view. Here we will put forward an alternative scheme which, although possibly equivalent at a computational level, may have a very different implementation. Our proposal is that when sample images of the specific object under the relevant transformations are not available, the system may generate virtual views of that object, using *image-based* transformations which are characteristic of the corresponding class of objects (Poggio and Vetter, 1992). We propose that the system learns these transformations from prototypical example views of other objects of the same class, with no need for 3D models. The idea is simple but it is not obviously clear that it will work. We will provide later a plausibility argument.

The problem of achieving invariance to image plane transformations such as translation, rotation and scaling, given only one model view, is also difficult, particularly in terms of biologically plausible implementations. But given a single model view, it is certainly possible to generate virtual examples for appropriate image-plane translations, scalings and rotations *without specific knowledge about the object*. This is not the case for the non-image-plane transformations we will consider here, caused by, for example, changes in viewpoint, illumination, facial expression, or physical attitude of a flexible or articulated object such as a body.

Within the *virtual views* theory, there are two extreme ways in which virtual views may be used to ensure invariance under non-image-plane transformations. The first one is to precompute all possible "virtual" views of the object or the object class under the desired group of transformations and to use them to train a classifier network such as the one of figure 1. The second approach – equivalent from the point of view of information processing – is instead to apply all the relevant transformations to the input image and to attempt to match the transformed image to the data base, which under our starting assumption, may contain only one view per object. These two general strategies may exist in several different variations and can also be mixed in various ways.

4.1 An example

Consider as an example of the general recognition strategy we propose the following architecture for biological face recognition based on our own work on artificial face recognition systems (Brunelli and Poggio, 1993; Beymer, 1993; see also Gilbert and Yang, 1993).

First the face has to be localized within the image and segregated from other objects. This stage might be template-based, and may be equivalent to the use of a network like that in Figure 3, with units tuned to the various low-resolution images a face may produce. From the biological point of view, the network might be realized by the use of low-resolution face detection cells at each location in the visual field (with each location examined at a resolution dictated by the cortical map, in which the fovea of course dominates), or by connections from each location in, say, V1 to "centered" templates (or the equivalent networks) in IT, or by a routing mechanism to achieve the same result with fewer connections (see Olshausen et al., 1992). Of course the detection may be based on disjunction of face components rather than on their conjunction in a full face template.

The second step in our face recognizer is to normalize the image with respect to translation, scale and image rotation. This is achieved by finding two anchor points, such as the eyes, again with a template-based strategy, equivalent to a network of the type of Figure 1 in which the centers are many templates of eyes of different types in different poses and expressions. A similar strategy may be followed by biological systems both for faces and other classes of objects. The existence of two stages would suggest that there are modules dedicated to detect certain classes of complex features – such as eyes – and other modules that use the result to normalize the image appropriately. Again there could be eye detection networks at each location in the visual field or a routing of relevant parts of the image – selected through segmentation operations – to a central representation in IT.

The third step in our face recognizer is to match the localized, normalized face to a data base of individual faces while at the same time providing for view-, expression- and illumination-invariance. If the data base contains several views of each particular face, the system may simply compare the normalized image to each item there (Beymer, 1993): this is equivalent to classifying the image using the network of Figure 1, one for each person. But if the data base contains only a single model view for each face, which is the problem we consider here, virtual examples of the face may be generated using transformations – to other poses and expressions – learned from examples of other faces (see Beymer, Shashua and Poggio, 1993; Poggio and Vetter, 1992; Poggio and Brunelli, 1992). Then the same approach as for a multi-example data base may be followed, but in this case most of the centers will correspond to "virtual examples".

4.2 Transformations and Virtual Examples

In summary, our proposal is to achieve invariance to non-image- plane transformations by using a sufficient number of views of the specific objects for various transformation parameters. If real views are available they should be used directly; if not, virtual views can be generated from the real one(s) using image-based transformations learned from example views of objects of the same class.

4.2.1 Transformation Networks

How can we learn class-specific transformations from prototypical examples? There are several simple technical solutions to this problem, as discussed by Poggio (1991), Poggio and Brunelli (1992) and Poggio and Vetter (1992). The proposed schemes can "learn" approximate 3D geometry and underlying physics for a *sufficiently restricted* class of objects – a *nice class*.[3] We define informally here *nice classes* of objects as sets of objects with sufficiently similar transformation properties. A class of object is *nice* with respect to one or more transformations. Faces are a nice class under viewpoint transformations because they typically have a similar 3D structure. The paperclip objects used by Poggio and Edelman (1990), Bülthoff and Edelman (1992 and in press) and by Logothetis and Pauls (in press) are *not nice* under viewpoint transformation because their global 3D structures are different from each other. Poggio and Vetter describe a special set of nice classes of objects – "linear classes" . For linear classes, linear networks can learn appropriate transformations from a set of prototypical examples. Figure 4 shows how by Beymer, Shashua and Poggio (1993) used the even simpler technique (linear additive) of Poggio and Brunelli (1992) for learning transformations due to face rotation and change of expression.

[3]The linear classes definition of Poggio and Vetter(1992) may be satisfactory, even if not exact, in a number of practically interesting situations such as viewpoint invariance and lighting invariance for faces.

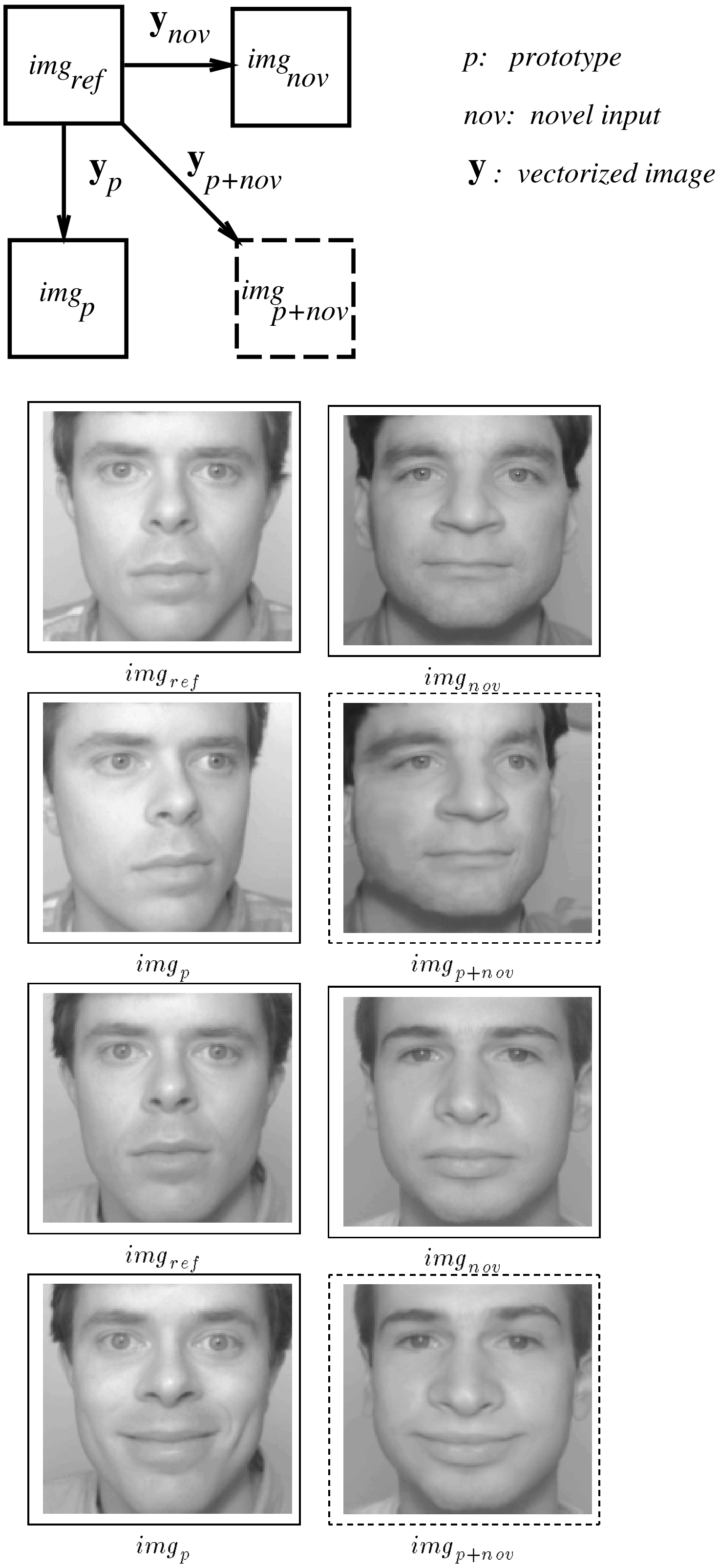

Figure 4: *A face transformation is "learned" from a prototypical example transformation. Here, face rotation and smiling transformations are represented by prototypes,* $\mathbf{y}_p$. $\mathbf{y}_p$ *is mapped unto the new face image* img_{nov}. *The virtual image* img_{p+nov} *is synthesized by the system. In a biological implementation cell activities instead than grey-levels would be the inputs and the outputs of the transformation. From Beymer, Shashua and Poggio, 1993*

In any case, a sufficient number of prototype transformations – which may involve shape, color, texture, shading and other image attributes by using the appropriate features in the vectorized representation of images – should allow the generation of more than one virtual view from a single "real" view. The resulting set of *virtual* examples can then be used to train a classification network. The argument so far is purely on the computational level and is supported only by preliminary and partial experiments. It is totally unclear at this point how IT cortex may use similar strategies based on learning class-specific prototypical transformations. The alternative model in which virtual examples are not explicitly generated and instead view-invariant features are learned is also attractive. Since networks such as Multilayer Perceptrons and HyperBasis Function networks may "find" some view-invariant features the two approaches may actually be used simultaneously.

4.3 An Alternative Visualization Route?

As we hinted earlier, an alternative implementation of the same approach to invariant recognition from a single model view is to transform the (normalized) input image using the learned transformations and compare each one of the resulting virtual views to the available real views (in this case only one per specific object). As pointed out by Ullman (1991), the cortex may perform the required search by generating simultaneously transformations of both the input image and the model views until a match is found.

The number of transformations to be tested may be reduced by first estimating the approximate pose and expression parameters of the input image. The estimate may be provided by a RBF-like network of the "analysis" type in which the centers are generic face prototypes (or face parts) spanning different poses, expressions and possibly illuminations[4]. They can be used if trained appropriately to do the *analysis* task of estimating state parameters associated with the image of the object such as its pose in space, its expression (if a face), its illumination etc. (see Poggio and Edelman, 1990; Beymer, Shashua and Poggio, 1993).

The corresponding transformation will then be performed by networks (linear or of a more general type).[5] Analysis-type networks may help reduce dramatically the number of transformations to be tried before successful recognition is achieved. A particular version of the idea is the following.

Assume that the data base consists of single views of different, say, faces in a "zero" pose. Then in the visualization route the analysis network provides an estimate of "pose" parameters; a synthesis network (Poggio and Brunelli, 1992; Librande, 1992; Beymer, Shashua and Poggio, 1993) generates the corresponding view of a prototype; the transformation from the latter prototype view to the reference view of the prototype is computed and applied to the input array to obtain its "zero" view; finally this corrected input view is compared with the data base of single views. Of course the inverse transformation could be applied to each of the views in the data base, instead of applying the direct transformation to the input image. We prefer the former strategy because of computational considerations but mixtures of both strategies may be suitable in certain situations.

This estimation-transformation route (which may also be called analysis-synthesis) leads to an approach to recognition in which parameters are estimated from the input image, then used to "undo" the deformation of the input image and "visualize" the result, which is then compared to the data base of reference views. A "visualization" approach of this type can be naturally embedded in an iterative or feedback scheme in which discrepancies between the visualized estimate and the input image drives further cycles of analysis-synthesis and comparison (see Mumford, 1992). It may also be relevant in explaining a role in mental "imagery" of the neurons in IT (see Sakai and Miyashita, 1991).

A few remarks follow:

1. Transformation parameters may be estimated from images of objects of a class; some degree of view invariance may therefore be achievable for new objects of a known class (such as faces or bilaterally symmetric objects (see Poggio and Vetter, 1992)). This should be impossible for unique objects for which prior class knowledge may not be used (such as the paperclip objects, Bülthoff and Edelman, 1992).
2. From the computational point of view it is possible that a "coarse" 3D model – rather like a marionette – could be used successfully to compute various transformations typical for a certain class of objects (such as faces) to control 2D representations of the type described earlier for each specific object. Biologically, this coarse 3D model may be implemented in terms of learned transformations characteristic for the class.
3. We believe that the classification approach – the one summarized by figures 1, 3, as opposed to the visualization approach – is the main route to recognition, which should be used with real example views when a sufficient number of training views is available. Notice that this approach is memory-based and in the extreme case of many training views should be very similar to a look-up table. When only one or very few views of the specific object are available, the classification approach may still suffice, if either a) view-invariant features are discovered and then used or b) virtual examples generated by the transformation approach are exploited. But this is possible only for objects belonging to a familiar class (such as faces). The analysis-synthesis route may be an additional, secondary strategy to deal with only one or very few real model views [6].

[4] Invariance to illumination can be *in part* achieved by appropriate preprocessing

[5] Of course in all of the modules described above the centers may be parts of the face rather than the full face.

[6] It turns out that the RBF-like classification scheme and its implementation in terms of view-centered units is quite different from the linear combination scheme of Ullman and

4. We have assumed here a supervised learning framework. Unsupervised learning may not be of real biological interest because various natural cues (object constancy, sensorimotor cues etc.) usually provide the equivalent of supervised learning. Unsupervised learning may be achieved by using either a bootstrap approach (see Poggio, Edelman and Fahle 1992) or an appropriate cost-functional for learning or special network architectures.

5 Critical predictions and experimental data

In this section we list a few points that may lead to interesting experiments both in psychophysics and physiology.

Predictions:

- **Viewer-centered and object-centered cells**. Our model (see the module of Figure 2) predicts the existence of viewer-centered cells (in the "hidden" layer) 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 2 is only a small part of an overall architecture. We expect therefore to find other types of cells, such as for instance pose-tuned, expression-tuned and illumination-tuned cells. Very recently N. Logothetis and Pauls (in press) have succeeded in training monkeys to the same objects used in human psychophysics and in reproducing the key results of Bülthoff and Edelman (1992). As we mentioned above, 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 conjecture (though 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 it would be sufficient to expose the monkeys to the objects without actually training them to respond in specific ways.

- **Cells tuned to full views and cells tuned to parts**. As we mentioned, we expect to find high-dimensional as well as low-dimensional centers, 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).

  Computationally, 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. 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 occlusions and the much lower requirements on system connectivity and complexity.

- **View-invariant features.** For many objects and recognition tasks there may exist features that are invariant at least to some extent (colour is an extreme example). One would expect this situation to occur especially in basic-level recognition tasks (but not only). In this case networks with one or very few centers and hidden units – each one being invariant – may suffice. One or very few model views may suffice.

- **Generalization from a single view for "nice" and "not nice" object classes**. An example of a recognition field measured psychophysically for an asymmetric object of a "not nice" class after training with a single view is shown in figure 5. As predicted from the model (see Poggio and Edelman, 1990), the shape of the surface of the recognition errors is bell-shaped and is centered on the training view. If the object belongs to a familiar and "nice" class of objects – such as faces – then generalization from a single view is expected to be better and broader because information equivalent to additional virtual example views can be generated from familiar examples of other objects of the same class. Ullman, Moses and Edelman (1993) report evidence consistent with this view. They use two "nice" classes of objects, one familiar – upright faces – and one unfamiliar – inverted faces. They find that generalization from a single training view over a range of viewpoint and illumination transformations is perfect for the familiar class and significantly worse for the unfamiliar inverted faces. They also report that generalization in the latter case improved with practice, as expected in our model.

  Notice again that instead of creating virtual views the system may discover features that are view invariant for the given class of objects and then use them.

- **Generalization for bilaterally symmetric objects**. Bilaterally symmetric objects – or objects that may seem bilaterally symmetric from a single view – are a special example of *nice* classes. They are expected from the theory (Poggio and Vetter, 1992) to have a generalization field with

---

Basri (1990). On the other hand a regularization network used for synthesis – in which the output is the image $\mathbf{y}$ – is similar to their linear combination scheme (though more general) because its output is always a linear combination of the example views (see Beymer, Poggio and Shashua, 1993).

additional peaks. The prediction is consistent with old and new psychophysical (Vetter, Poggio and Bülthoff, 1994) and physiological data (Logothetis and Pauls, in press).

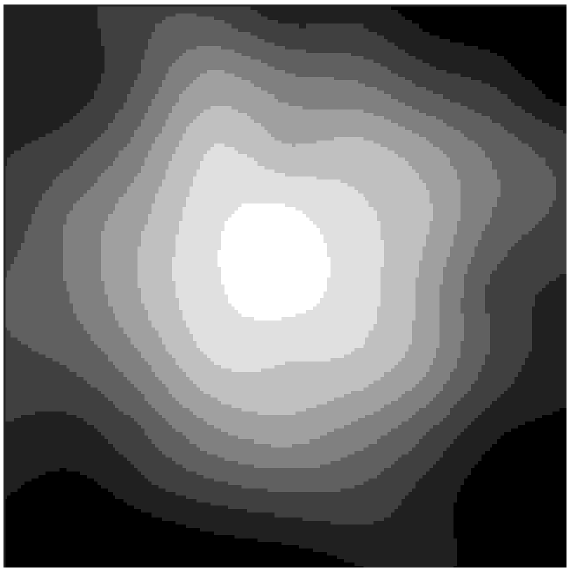

Figure 5: *The generalization field associated with a single training view. Whereas it is easy to distinguish between, say, tubular and amoeba-like 3D objects, irrespective of their orientation, the recognition error rate for specific objects within each of those two categories increases sharply with 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* $\pm 60°$ *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 (1992). As predicted by our model viewpoint independence can be achieved by familiarizing the subject with a sufficient number of real training views of the 3D object. For objects of a nice class the generalization field may be broader because of the possible availability of virtual views of sufficient quality.*

References

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

[2] D. Beymer, A. Shashua, and T. Poggio. Example based image anaysis and synthesis. A.I. Memo No. 1431, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[3] D. J. Beymer. Face recognition under varying pose. A.I. Memo No. 1461, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

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

[5] R. Brunelli and T. Poggio. Hyberbf networks for real object recognition. In *Proceedings IJCAI*, Sydney, Australia, 1991.

[6] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.

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

[8] H.H. Bulthoff 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.

[9] S. Edelman. Features of recognition. CS-TR 91-10, Weizmann Institute of Science, 1991.

[10] 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.

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

[12] J.M. Gilbert and W. Yang. A real-time face recognition system using custom vlsi hardware. In *IEEE Work on Computer Architectures for Machine Vision*, 1993.

[13] F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. A.I. Memo No. 1430, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[14] S. Librande. Example-based character drawing. Master's thesis, M.S., Media Arts and Science Section, School of Architecture and Planning, Massachusetts Institute of Technology, Cambridge, MA, September 1992.

[15] N.K. Logothetis and J. Pauls. Psychophysiological and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, in press, 1994.

[16] N.K. Logothetis, J. Pauls, H.H. Bülthoff, and T. Poggio. View-dependent object recognition in the primate. *Current Biology*, in press, 1994.

[17] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976.

[18] B. W. Mel. The clustron: Toward a simple abstraction for a complex neuron. In S. Hanson J. Moody and R. Lippmann, editors, *Neural Information Processing Systems*. Morgan-Kaufmann, San Mateo, 1992.

[19] D. Mumford. The computational architecture of the neocortex. *Biological Cybernetics*, 1992.

[20] B. Olshausen, C. Anderson, and D. Van Essen. A neural model of visual attention and invariant pattern recognition. CNS Memo. 18, California Institute of Technology, Computation and Neural Systems Program, 1992.

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

[22] D. I. Perrett and M.W. Oram. The neurophysiology of shape processing. *Image and Vision Computing*, 11(6):317–333, 1993.

[23] 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.

[24] T. Poggio. 3D object recognition and prototypes: one 2D view may be sufficient. Technical Report 9107–02, I.R.S.T., Povo, Italy, July 1991.

[25] T. Poggio and R. Brunelli. A novel approach to graphics. A.I. Memo No. 1354, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

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

[27] 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.

[28] 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, 1990.

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

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

[31] T. Poggio and A. Hurlbert. Sparse observations on cortical mechanisms for object recognition and learning. A.I. Memo No. 1404, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[32] T. Poggio and A. Hurlbert. Sparse observations on cortical mechanisms for object recognition and learning. A.I. Memo No. 1404, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, in press, 1993.

[33] 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.

[34] K. Sakai and Y. Miyashita. Neural organization for the long-term memory of paired associates. *Nature*, 354:152–155, 1991.

[35] P.G. Schyns and H.H. Bülthoff. Conditions for viewpoint dependent face recognition. A.I. Memo No. 1432, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[36] A. Shashua. *Geometry and Photometry in 3D visual recognition.* PhD thesis, M.I.T Artificial Intelligence Laboratory, AI-TR-1401, November 1992.

[37] A. Shashua. Illumination and view position in 3D visual recognition. In S.J. Hanson J.E. Moody and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 404–411. San Mateo, CA: Morgan Kaufmann Publishers, 1992. Proceedings of the fourth annual conference NIPS, Dec. 1991, Denver, CO.

[38] K. Tanaka. Neuronal mechanisms of object recognition. *Science*, 262:685–688, 1993.

[39] T.Poggio and S.Edelman M. Fahle. Fast perceptual learning in visual hyperacuity. *Science*, 256:1018–1021, 1992.

[40] S. Ullman. Sequence-seeking and counter-streams: a model for information flow in the cortex. A.I. Memo No. 1311, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991.

[41] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1006, 1991.

[42] T. Vetter, T. Poggio, and H. Bülthoff. The importance of symmetry and virtual views in threedimensional object recognition. *Current Biology*, 4:18–23, 1994.

[43] S. Ullman Y. Moses and S.Edelman. Generalization across changes in illumination and viewing position in upright and inverted faces. In *Perseption Volume 22 Supplement ECVP 93 Abstracts*, page 25, 1993.