On the Representation of Novel Objects: Human Psychophysics, Monkey Physiology and Computational Models

by

Emanuela Bricolo

Laurea, Electrical Engineering
Universitá di Padova (1989)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational Neuroscience

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1996

Signature of Author .....
Department of Brain and Cognitive Sciences
June 27, 1996

Certified by .....
Dr. Tomaso Poggio
Uncas and Helen Whitaker Professor
Thesis Supervisor

Accepted by .....
Gerald E. Schneider
Chairman
Department Graduate Committee

On the Representation of Novel Objects: Human Psychophysics, Monkey Physiology and Computational Models

by

Emanuela Bricolo

Submitted to the Department of Brain and Cognitive Sciences
on June 27, 1996, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computational Neuroscience

Abstract

We recognize objects in three dimensions, despite the diverse two-dimensional projections produced by changes in orientation. To achieve recognition, visual input is compared to stored shapes. This stored models may take many forms. Much support has been given recently to the proposal that object representation is specific to viewpoint. This approach to the problem of object recognition assumes that stored models are encoded in particular orientations, usually those orientations from which the input shapes were first observed. This thesis combines psychophysical and physiological experiments, together with computational simulations, providing further support for this theory. These viewpoint-dependent representations are characterized in detail for the case of a particular class of novel wire-like three-dimensional objects.

In Part I, in accordance with viewpoint-specific theories, the psychophysical results reveal that performance is consistently viewpoint-dependent, and systematically disrupted by rotation in depth, more so than by deformation of the object itself. Dependence of performance on object deformation is studied in detail, and it is suggested that both qualitative and quantitative information is used in shape representation.

In Part II, neurophysiological data show that cells in inferotemporal cortex display properties suggesting a representation of objects as a collection of views, each coded by one or more neurons. The recordings reveal a small population of neurons with remarkable selectivity for individual views of a set of objects which monkeys learned to recognize. An analysis of population responses to different views of objects provides further evidence that neural representation of object shape depends on abstract two-dimensional views.

Part III describes a model of these view-tuned units. The approach consists of representing a view in terms of a few local features, computed and stored during the training phase. Each feature is represented as the set of responses of oriented filters at one location in the image. During recognition, the system computes a robust conjunction of the best matches to the stored features. Simulations show that the

model is consistent with physiological data from single-cell responses.

Thesis Supervisor: Dr. Tomaso Poggio
Title: Uncas and Helen Whitaker Professor

Acknowledgments

Foremost I am indebted to my thesis advisor, Tomaso Poggio, for providing support, guidance and encouragement and for putting up with a traveling graduate student. A special thanks to Heinrich Bülthoff, Nikos Logothetis and Michael Jordan not only for serving on my thesis committee but for supporting me in many ways during my years as graduate student.

Many people have shared my life during these years and gave me many nice memories to take home with me John Houde, Daphne Bavalier, Janine Mendola and in particular the "italian connection" Alessandra Angelucci, Francesca Gandolfo, Eleonora Fusco, Domenico Seminara, Mark Carter, Luca Griguolo, Lamberto e Elena Piron, and Federico Girosi.

A special thanks goes to Jan Ellertsen for being so helpful and understanding. And I will be always grateful to everybody at CBCL who helped me with big and little problems. Thanks to Marney Smyth whose contribution to this thesis was invaluable, to MaryPat Fitzgerald for taking such good care of me, to Gadi Geiger for being a great officemate. A big thanks to Leonardo for his help and for his great patience in getting my computer problems fixed. I would also like to thank my collaborators Janine Mendola, Josh Tenenbaum and Jon Pauls.

Also a big thanks to all the people that have made my stay outside Boston enjoyable Stefan Treue, Jon Pauls, David Sheinberg, David Leopold, Yanna Logothetis, Thomas Vetter and Pascal Mamassian.

Last I want to thank my parents for being near me during my interminable "student's" years. This thesis is dedicated to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We recognize visually-presented objects quickly and effortlessly. Our visual system can perform this task despite substantial changes in the retinal images. If an object maintains a rigid structure over time, then as the observer-relative viewpoint varies, the outline, size, and position of the object retinal projections can be significantly modified. The pattern of image luminance intensities also changes, due to alterations in illumination conditions (intensities and position of light sources). Therefore, for reliable recognition, we need to extract and store a description of objects which is invariant to image transformations, that is to image size, scene illumination, and object position in the visual field. On the other hand, the representation needs also to be detailed, in order to allow us to discriminate between similar objects, or more to the point between different objects that cast similar images. More specifically, in the case of rigid objects, small differences in three-dimensional (3D) shape should be correctly represented as different objects, while even bigger differences in two-dimensional (2D) projection should be correctly interpreted as images of the same object. But which representation will satisfy these requirements? What kind of information is extracted from the image and coded?

Very different processes lead to the recognition of visually-presented objects. Not only shape, but also color, texture, and motion information, could be used alone for correct identification (Ullman, 1989). Nonetheless, this work will mainly focus on shape-based recognition since shape is the most common cue, the one we mostly rely

on for the purpose of recognition.

Recognition is not a well defined term, objects can be recognized at different hierarchical levels as humans organize information in equivalence classes. Each class or category groups items with common functional or perceptual properties, also called common features, or *cues* (Rosch et al., 1976). The ability of a given feature to predict class membership is called cue validity, and the cue validity of a category is the sum of the cue validity of all the cues that define that category. We can distinguish three hierarchically organized categorical levels on the base of the number and validity of their defining cues. The categories with highest cue validity are called *basic-level* categories, e.g. cat or dog. More specific categories, those that require detailed shape-base distinctions, e.g. Siamese or Doberman, are called *subordinate-level* categories, while more general ones, e.g. mammals, are called *superordinate-level* categories (Rosch et al., 1976). Neuropsychological results report that recognition at different categorization levels may involve different neural circuitry (Damasio, 1990; Damasio et al., 1990).

Particularly at the basic and subordinate levels, object categorization is deemed to rely principally on perceptual cues (Logothetis and Sheinberg, 1996), and there is a clear implication that conceptual categories (at these levels) reflect more than just linguistic constructs. Therefore, in our experiments we will test subjects discrimination performance with tasks that require subordinate level classification. This performance has implications for the *form* of the internal representation which the visual system constructs in response to pattern stimuli (Marr, 1982). This property allows us to gain insight on the representation used by the visual system, while identifying the cases in which recognition performance breaks down more easily.

The issue of segmentation will not be considered. This is a fascinating problem that may need to be solved before the representation is extracted. Instead, in the experimental paradigm *isolated objects against uniform background* will be used, since understanding the neural representation of objects is a very complex and difficult task, even when very simple objects viewed in isolation are considered.

1.1 Theories of object recognition

In the area of three-dimensional (3D) object recognition, the current debate focuses on the choice of the reference frame in which the objects are described. There are two main schools of thought. The first favors the idea of object-centered 3D representations, where an observed object is assigned the same representation regardless of its orientation, size, or location (Marr and Nishihara, 1978; Biederman, 1987; Corballis, 1988). The second school favors viewer-centered representation (Rock and DiVita, 1987; Bülthoff and Edelman, 1992; Tarr and Pinker, 1989), where an observed object has multiple representations depending on its orientation with respect to the viewer.

The former approach implies a heavy load on computation: The structure of the object must be extracted from each image. On the other hand, the alternative approach relies heavily on memory. In the extreme case, no preprocessing needs to be done at all, and a collection of views of the same object under all possible different conditions is stored into memory under the same label. This approach is usually dismissed as unfeasible, because of the combinatorial explosion in memory requirements it imposes on the system.

A wide range of models can be found between this two extremes. For example, classifiers that use simple features like oriented lines, and angles between line segments, can obtain recognition moderately invariant to rotation both in and outside the image plane (Edelman and Poggio, 1992; Duda and Hart, 1973; Cavanagh, 1978). Lines are often chosen, since they seem to be encoded in the early stages of visual processing (Hubel and Wiesel, 1959). Angles between lines are for example a feature extraction procedure that gives complete rotation invariance in the image plane. The memory load is reduced, since in the case of feature extraction, there is a small generalization for each view. The number of exemplars needed depends on the algorithms used to compute membership to a class.

These different models would influence the choice of algorithm used in the computation, and would also utilize different representations (Marr, 1982).

1.2 Overview of the thesis

In summary, this thesis addresses the issue of *characterizing the representation of simple, isolated, rigid objects for shape-based recognition.* To address this issue exhaustively, the problem is analyzed merging information from various fields of vision research — specifically psychophysics, neurophysiology, and computational modeling.

1.2.1 Human Psychophysics

Objects can be recognized by their projected shapes alone, even if these projected shapes vary greatly as the observed viewpoint changes. Current research is focused on whether this recognition is accomplished by matching an input shape to an orientation-dependent or an orientation-independent representation.

Part I addresses the issue of view-dependent representation, and whether this representation is systematically disrupted more by rotation in depth than by deforming the object image. A number of experiments tested recognition performance when objects have to be discriminated among other objects with varying degrees of similarity. The main findings can be summarized as follows:

1. Experiment 1, tests recognition performance for a number of views of objects against a group of distractors with varying degree of similarity. At low noise level, i.e. when objects are very similar, subjects are unable to discriminate, suggesting that the view-tuning is obtained by storing a prototype that allows a certain amount of noise in the description. As the similarity decreases, the recognition becomes easier but differentially so for various views of the object. This confirms that the frame of reference used in subordinate level classification is viewpoint dependent.

2. Subsequent experiments (Experiment 2 to Experiment 5) test recognition of a single view of an object in more detail. The results suggest that, at least for the class of objects used in our experiments, this representation may be strongly configurational, weighting some characteristic parts of the image more

than others.

3. A final set of experiments tested the dependence of object representation on object-dependent transformation and showed that the representation is indeed size- and position-dependent.

1.2.2 Monkey Physiology

Part II provides evidence for a similar view-dependency of recognition in the non-human primate. Moreover, combined psychophysical and electrophysiological experiments, tested monkeys for the ability to generalize recognition for views generated by rotating objects around any arbitrary axis. Throughout this process, the activity of single units in inferotemporal cortex was recorded. These experiments were performed in collaboration with Jon Pauls and Nikos Logothetis at Baylor College of Medicine.

Our results provide evidence supporting a viewer-centered representation of objects in the primate, at least for subordinate level classifications. The main findings of this study can be summarized as follows:

1. Even when complete information about the structure of an object is available to the subject, recognition at the subordinate level depends on the object's attitude, both for monkeys as well as for human subjects.

2. A memory-based, viewer-centered recognition system is not an implausible mechanism for object-constancy. Both theoretical work, and the results presented here, suggest that only a small number of object views need to be stored in order to achieve perceptual invariance.

3. A small population of IT neurons has been found to respond selectively to individual members of the object-classes tested in this study. The response of some neurons is a function of the object's view. The discharge rate of many IT neurons is found to be a bell-shaped function of orientation centered on

a preferred view and view-tuning is observed only for those views that the monkeys can recognize.

4. A subset of the view-selective units were also tested for position and size changes of the preferred view. In our experiments, the responses of IT neurons lie on a continuum from cells invariant to either scale, position, or both, to cells with varying degrees of sensitivity to these parameters.

1.2.3 Computational models

Part III describes a recognition architecture that could underlie the view-dependent recognition performance of humans and monkeys reported earlier. This architecture relies on small-scale networks with units that are broadly tuned to views of a learned object – similar to the neurons recorded in IT. Recent computer simulations have shown that a simple network can recognize $3D$ objects by *interpolating* between a small number of stored views (Poggio and Edelman, 1990). This network uses a small set of sparse data, corresponding to an object's training views, in order to synthesize an approximation of a multivariate function (Poggio and Girosi, 1990b) representing the object. The initial simulations were done using wire-like computer-rendered objects, whereby a view was represented either by the ordered pairs of x, y coordinates of the wire vertices (Poggio and Edelman, 1990), or by their image-plane segment-orientations (Logothetis et al., 1994). However, the model does not specify the inputs to the view-tuned units and their internal organization. The units are not necessarily tuned to a complete view of the object, high level complex features could be used instead. The proposed model of these view-tuned units is consistent with physiological data from single cell responses. The approach consists of representing a view in terms of a few local features, computed by V1-like cells, which can be regarded as local configurations of grey-levels. Though the model is still far from being a complete neuronal model, it is already being used to make useful predictions for physiological experiments which are currently underway. The main predictions are:

1. The model units are view-selective. Their behavior is comparable to that of view-tuned cells as characterized by their response to rotation in depth of the objects they are tuned to and by the relative response to distractors.

2. The model also predicts the complex behavior seen for partly occluded objects. An occlusion can or cannot disrupt the response, depending on the number and position of the extracted features.

Part I

Human Psychophysics

Chapter 2

Object specific transformations

2.1 Introduction

Shape is the most important cue to recognition, even if by no means the only one. Objects can be recognized by their projected shapes alone, even if these projected shapes vary greatly as the observed viewpoint changes. Current research is focused on whether this recognition is accomplished by matching an input shape to an orientation-dependent representation. A close examination of experimental results reveals that both the class of objects and the paradigm used in the experiments greatly influence the findings. When enough time was given to the subjects (reaction times of the order of seconds) and simple stimuli were used, mental rotation experiments could be interpreted as a renormalization of the object to a "standard" view before its comparison with the memorized version (Shepard and Metzler, 1971).

Experiments that use slightly more complex objects, and calling for fast recognition rates, suggest that the human visual system recognizes 3D objects by 2D view interpolation, using an image based representation scheme (Bülthoff and Edelman, 1992; Rock et al., 1989; Tarr and Pinker, 1989). It was shown that subjects' error rates grew with increasing distance from the view used in training. The objects were novel to the subjects that had to commit the target to memory so as to perform the task. All these combined results suggest that recognition at the subordinate level depends on the object attitude, even when complete information about the structure

of the object is available to the subject.

2.2 Experimental Paradigms

All the experiments described in this thesis employ one of two tasks in which subjects were asked to explicitly remember whether a given object had been previously displayed. In the case of both tasks, we test generalization to novel aspects of a given object. The object on which the tests are based is called the *target.* Usually, it is first presented to the subject in a specific and controlled attitude for a certain amount of time, so that the subject can memorize it. In a subsequent phase, one or many *distractors* (objects similar but not identical to the target) are introduced. The subject's ability to discriminate between them and the target is then tested, where one or more of the target parameters (depending on the experiment) may be manipulated.

Such judgments are usually referred to as *explicit* memory tasks in comparison to *implicit* memory tasks such as naming (Schacter, 1987). Explicit memory tasks have been deemed inappropriate for testing representation since they often find significant effects where implicit tasks fail to. This dissociation is still debated, especially in the case of viewpoint effects (Cooper and Schacter, 1992; Tarr and Pinker, 1989).

2.2.1 Sequential matching paradigm

In the **sequential matching** paradigm, two objects are shown in each trial. A fixation spot appears initially, for a specific time t_F, then the target is presented for a set time t_1, so as to allow the subject to study it. A delay t_D period follows, during which the fixation spot alone is presented.

When t_D has elapsed, a second object is displayed and the subject's task is to report if the presented objects are the same or different. In a *same* trial, the objects have the same 3D shape even if there may be a change in the viewing parameters of one object with respect to those of the other. In a *different* trial, the objects' 3D shapes differ. The subject records a response by pressing one of two appropriately-labeled keys on the keyboard. There are a number of variations of this paradigm. For

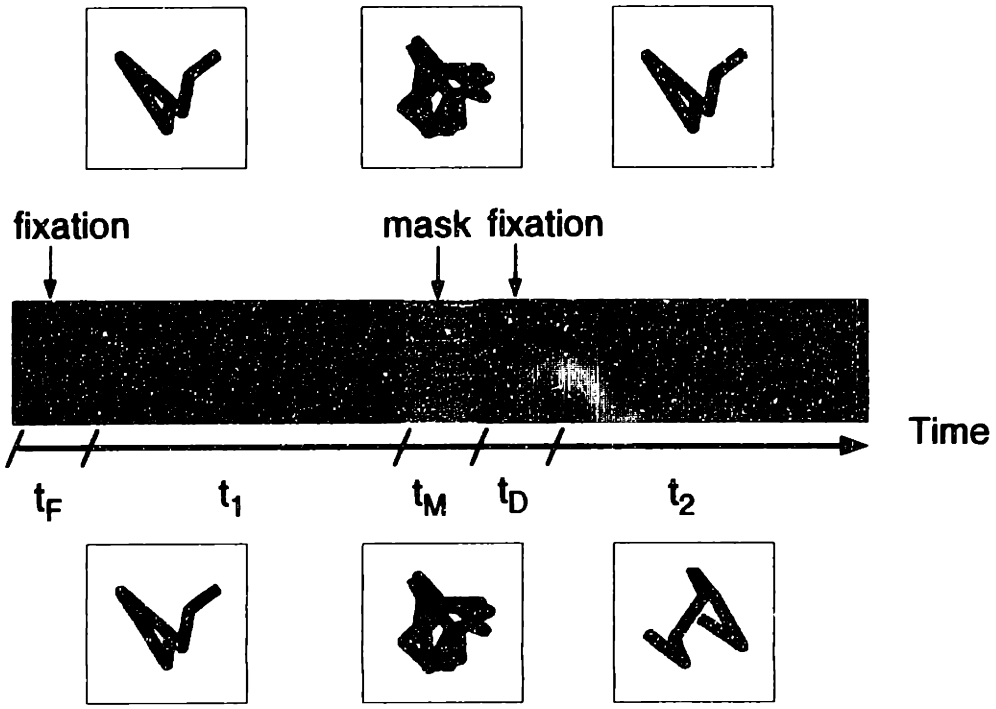

Figure 2-1: Sequential matching paradigm. Top: example of *same* trial. Bottom: Example of *different* trial.

instance, the fixation periods could be substituted with blank screen presentations. Another variation, used in our studies, consists of the introduction of another object, called *mask* during the delay period t_D, see Figure 2-1. The mask is introduced in order to terminate or interfere with visual processing of the target.

2.2.2 Match-to-sample paradigm

The **match-to-sample** paradigm consists of two parts: a preliminary study phase followed by a test phase. In the study phase, the fixation spot initially appears for a specific time, t_F, then the target is presented for a certain time, t_1, in order for the subject to study it (the target is also called *study object*). There is usually no presentation of a mask.

The test phase commences when a delay time, t_D, has elapsed. The test phase here comprises of a sequence of tests on multiple objects, with delay phases (t_F) interspersed between each of the *test objects*. Each test object – in our experiment usually between 4 and 10 – presented in the sequence appears for time t_2, and may

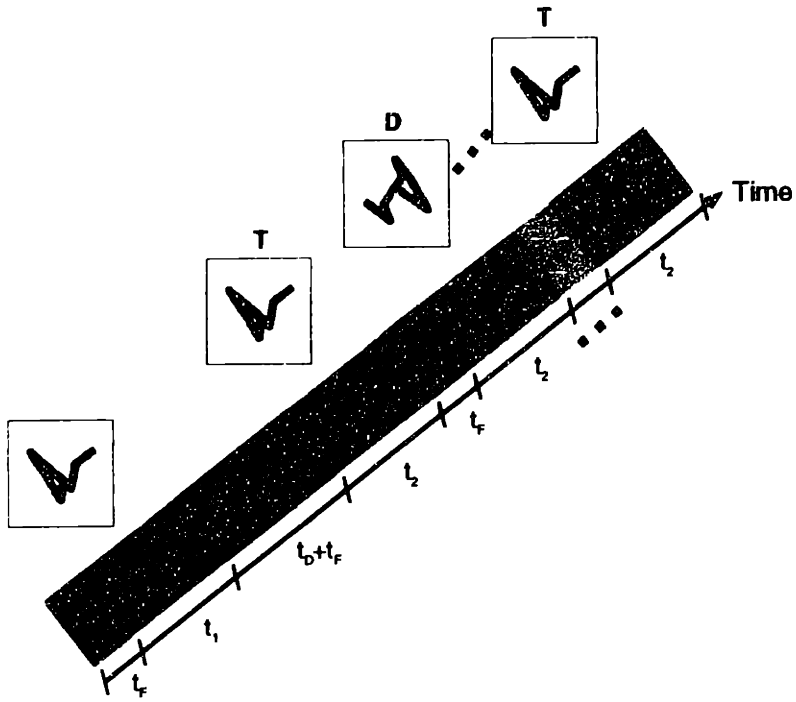

Figure 2-2: Match to sample paradigm. Above the sample figures are shown appropriate responses (T=target, D=distractor).

be either the target or one of a number of distractors, see Figure 2-2. As each object appears, the subject records a target (T) or distractor (D) decision by pressing one of two appropriately labeled keys on the keyboard. The next sequence of study/test phase starts with a new study object.

2.3 Experimental results

In the following experiments, we are addressing the issue of view-dependent representation and of the type of information stored in the representation, and whether this representation is systematically disrupted more by rotation in depth than by deforming the object image.

2.3.1 Experiment 1

Previous experiments with bent wire-like objects (Bülthoff and Edelman, 1992) have shown that recognition is highly sensitive to rotation in depth. The following exper-

iment addresses a similar question. We measure the range of generalization from a single viewpoint in the case where novel objects are learned in the presence of all possible cues to recognition. We are also using a controlled set of distractors in order to define the specificity of the representation.

Method

Subjects Thirteen paid subjects, all students at the University of Tübingen, Germany, participated in the experiment. All had normal or corrected-to-normal vision. All subjects were unaware of the purpose of the experiment.

Apparatus The stimuli were generated on an SGI workstation and were displayed on a CRT monitor. The screen was positioned at a distance of 114 cm from the subjects and the resolution was 1280 pixels (340mm) wide by 1024 pixels (272 mm) high. The stimuli were viewed in stereo with the aid of LCD shutter glasses of Crystal Eye Eyewear. Subjects were asked to use a chin-rest in order to limit head movements and to maintain a fixed viewing distance from the screen. A uniform black background field was employed: it subtended 13.4° x 16.6° of the visual field. At this distance the resolution was 77 pixels/degree.

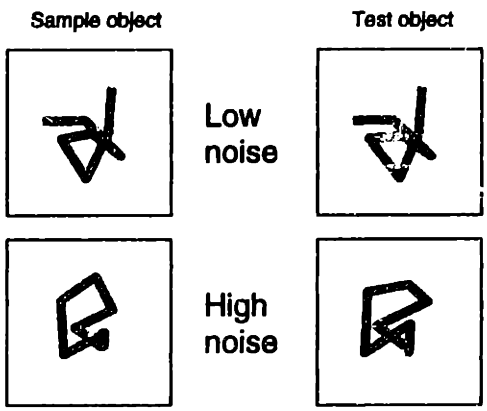

Figure 2-3: *Experiment 1:* Examples of stimuli with corresponding noisy distractors.

Stimuli The stimuli were gray level images obtained as perspective projections of shaded 3D objects. A set of 600 wire-like objects to be used as targets was created from Class II defined in Appendix A. Since we wanted to manipulate object similarity, three distractors were generated for each target. The distractors were created by

adding a fixed amount of noise to the vertices of the corresponding target object. Thus, when creating a distractor, the three-dimensional distance from the position of a given vertex in the target object, and the position of the corresponding vertex in the distractor object was kept constant, while the direction of the noise was randomly picked. Three noise levels were used (respectively 7.5%, 15% and 22.5% of average segment length). Examples of these objects and the corresponding distractors are shown in Figure 2-3 for the case of low and high noise level. Because the test object obtained via this manipulation did not satisfy the constraint of equal segment length, this constraint was not maintained when generating the study objects. Objects were rendered to be approximately 0.9° in diameter. A mask object was used in the test. It was composed of 20 tubular segments serially connected without any requirements on angle or intersection constraints.

Procedure Prior to the actual experiment, the subjects were tested for stereo vision with the use of random dot stereograms. An equilateral triangle was presented at different disparities, both crossed and uncrossed, spanning the range later used in the testing. The triangle was invisible when viewed monocularly. The subjects had to judge whether the triangle was pointing upwards or downwards, and consequently press one of two appropriately-labeled keys. The stereo test data were analyzed so to exclude any subjects with possible stereo deficits. All the tested subjects had no stereo deficit, and they could classify correctly ($> 95\%$ correct) the triangles at both crossed and uncrossed disparities.

The subjects were allowed to become familiar with the type of stimuli that were going to be used in the testing session, via an interactive program. The program allowed the subjects to rotate a sample object around any axis. The sample object was not used in later tests.

In the actual experiment, objects were always presented foveally at a fixed distance from the viewer (no size change) and under fixed illumination conditions. A sequential matching paradigm was used (see Section 2.2.1). At each trial, the fixation point appeared on the screen for 500ms, then the first object was presented stereoscopically for 8s. The object was then replaced by the mask for 500ms. Immediately after the

mask disappeared, the second object was presented for up to 3s, or until the subject response, whichever was shorter. The subjects had to judge if the two presented objects were identical. After each response, the subject was given audio feedback on the performance. If the subject response took longer than 3s, the trial was aborted and repeated later to keep a completely balanced design. After every 25 study/test pairs, the subject was instructed on overall performance. There were four blocks, consisting of 150 study/test pairs. Each block lasted on average 30 minutes. Usually, subjects ran two successive blocks in one session. There was no practice set.

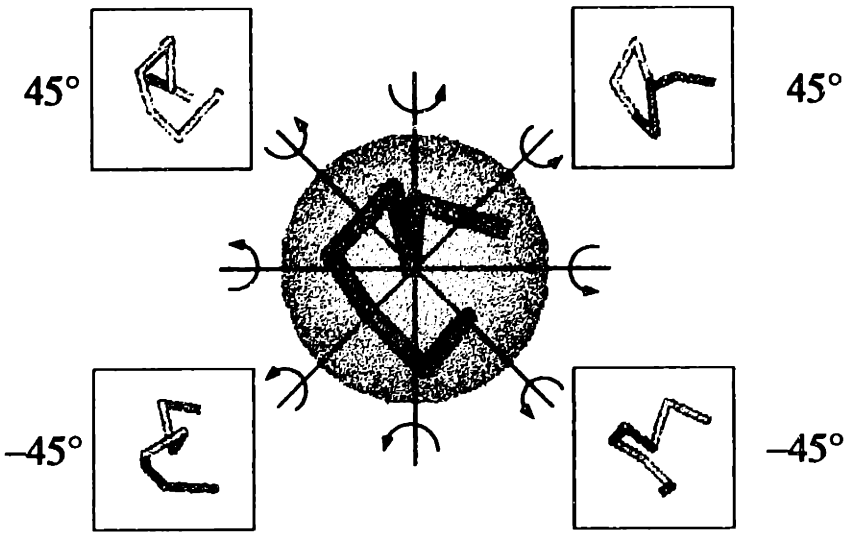

Figure 2-4: *Experiment 1:* A sample object is shown inside a sphere. The four axis around which the objects are rotated, are shown together with 4 examples of images generated by rotating the object. Degrees of rotation are shown next to the images

In a same trial, both objects were identical in shape; in different trials, the second object was the corresponding distractor. In both types of trials, the second object could appear rotated in depth. Twenty-five different viewpoints were tested. The zero view, the view at which the target was first presented, was picked randomly out of all the possible views. The other 24 views were obtained by rotating the objects $\pm 15°$, $\pm 30°$, and $\pm 45°$, around four equispaced axes going through their center of mass (see Figure 2-4). Each viewpoint was tested 10 times for each subject, for both targets and distractors, at each of the noise levels.

In each trial, a different object was tested. Different objects were tested for each noise level, and for each possible viewpoint, in order to reduce object dependent results. Tests with different levels of difficulty were intermixed. Target and distractor were tested with equal frequency.

Results

The hit rate – percentage of targets recognized correctly – was computed for each subject at each of the 25 levels of target orientation. Each hit rate was based on 4 trials. The mean hit rate was then computed, averaging across subjects. Since the noise level was also a variable of interest, three sets of false alarm rates where computed, one for each noise level. For each of these sets, a false alarm rate was computed for each subject, based also on four trials and then averaged across subjects.

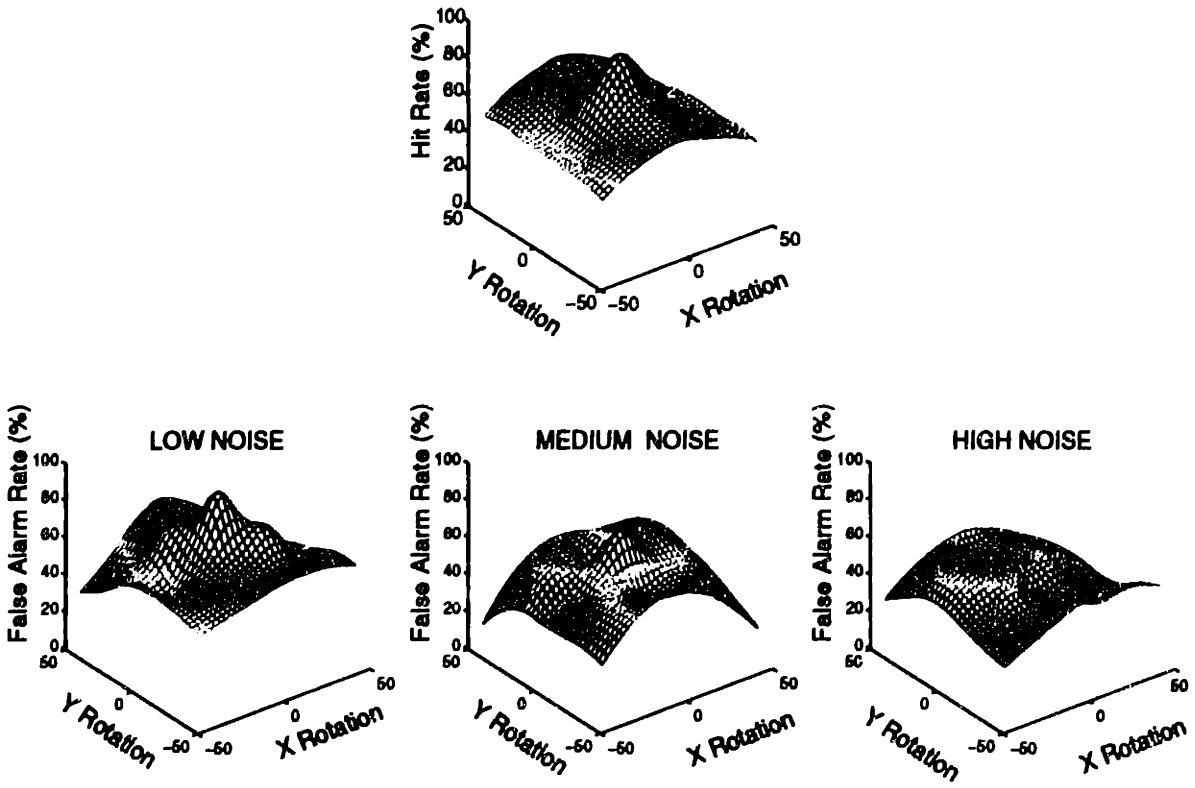

Figure 2-5: *Experiment 1.* TOP: Hit Rate for different viewpoints, values on are interpolated from the 25 tested viewpoints. BOTTOM: Interpolated False Alarm Rates for low (7.5%), intermediate (15.0%), and high (22.5%) noise levels.

In Figure 2-5, we show data dependency on viewpoint for targets and noisy dis-

tractors at different noise levels. The hit rates and false alarm rates computed as shown above were used to approximate the performance for viewpoints not tested in the experiment. The plot is obtained by a linear combination of normalized gaussians with the same sigma, each centered at one of the tested viewpoints.

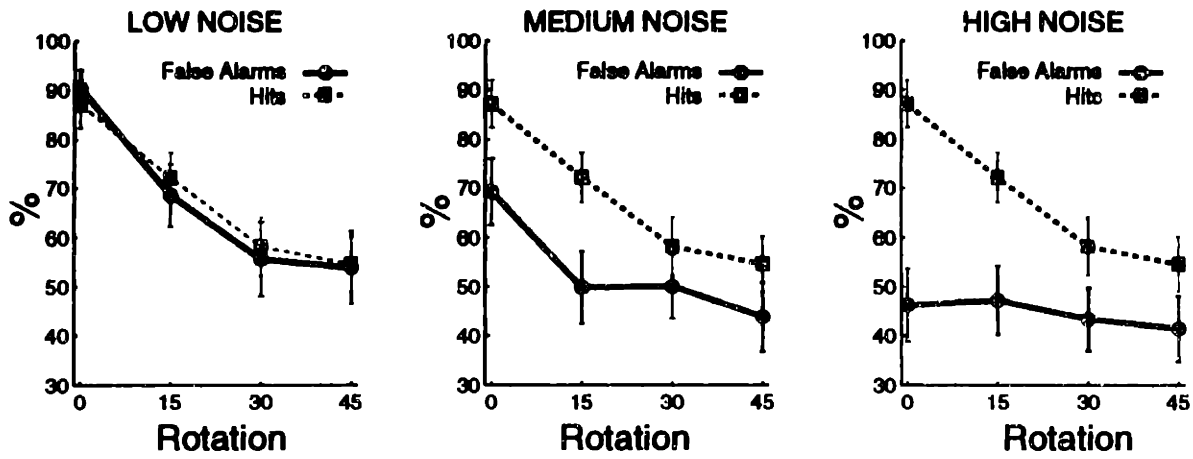

Figure 2-6: *Experiment 1:* The three graphs show hit rate (dashed) and false alarm rate for the three noise conditions. Note that the hit rate is the same in all the figures, because the testing was intermixed.

For further analysis, we collapsed the data across axis of rotation for each subject. The mean hit rate and the mean false alarm rates for each level of target orientation ($0°, 15°, 30°, and 45°$) are presented in Figure 2-6. The error bars indicate the standard error of the mean computed across the 13 subjects.

Since the noise level of the distractors was a variable of interest, false alarm rates were used as measure of recognition. A two-way Analysis of Variance (ANOVA) was conducted on the false alarm rates with Noise Level (No noise, Low, Intermediate, and High) and Distractor Orientation ($0°, 15°, 30°$, and $45°$) as within-subject variables. There is a main effect of both Noise Level ($F(2, 144) = 25.4, p < 0.0001$) and Orientation ($F(2, 144) = 14.1, p < 0.0001$) but also an significant interaction between the two variables ($F(2, 144) = 3.1p < 0.001$).

Discussion

The results confirm earlier rotation experiments: The recognition is sharply tuned to the zero view when the distractors are very different from the targets (Bülthoff

and Edelman, 1992; Tarr and Pinker, 1989; Rock and DiVita, 1987). At low noise level, subjects are unable to discriminate, suggesting that the view tuning is obtained by storing a prototype that allows a certain amount of noise in the description. As the similarity decreases the recognition becomes easier but differentially so for different views of the object strengthening the idea that the frame of reference used in subordinate level classification is viewpoint dependent.

These results are consistent with the view interpolation framework proposed by Poggio and Edelman (Edelman and Poggio, 1992). These networks predict the fall off in performance when the viewing angle is significantly different from the angles at which the object was learned.

2.3.2 Experiment 2

In this experiment, we address the question of which spatial dimensions characterize object recognition (2D v. 3D). From the previous experiment, we arrived at the conclusion that the frame of reference used in subordinate level classification is viewpoint-dependent. This does not specify the spatial dimension of these views. Some rough information about the third dimension could be extracted and used in the representation.

Method

Subjects Ten paid subjects participated in this experiment, all students at the Massachussetts Institute of Technology. All had normal or corrected-to-normal vision. All subjects were unaware of the purpose of the experiment.

Apparatus

The apparatus was similar to that described in Experiment 1, except that the distance between the subjects and the screen was of 109 cm, thus the background field subtended 14.0° x 17.3° of the visual field. The stimuli were viewed in stereo, with the aid of LCD shutter glasses of Crystal Eye Eyewear. Subjects were asked to use a chin-rest in order to limit head movements, and to maintain a fixed viewing

distance from the screen. At this distance, the resolution was 74 pixels/degree[1]

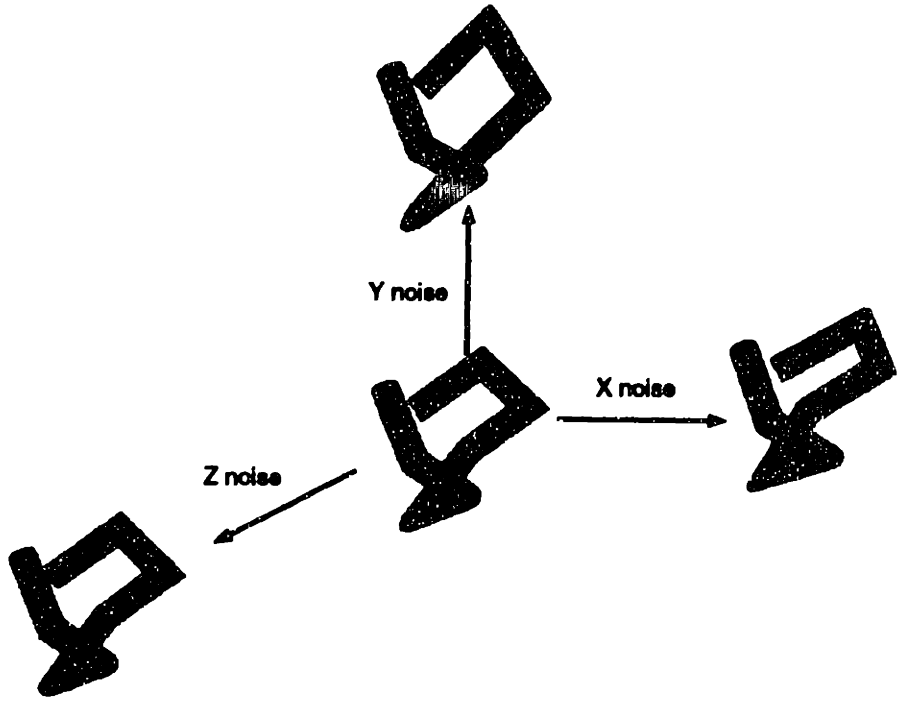

Figure 2-7: *Experiment 2:* Example of an object and corresponding noisy distractors.

Stimuli The stimuli were gray level images obtained as perspective projections of shaded 3D objects. A set of 40 wire-like objects was generated from Class II described in Appendix A. In this experiment, as in the previous one, the distractor set was the variable of interest. Each target was used as a prototype from which three distractors were generated. In this experiment, these three distractor had noise added respectively only to the x, y, and z coordinates of the target vertices (see Figure 2-7). The amount of noise — displacement from a given vertex — was identical for the three distractors of a given target, but different for different targets.

Procedure In this experiment a *match-to-sample* paradigm was used (see Section 2.2.2). The study object was presented for about 15s undergoing a motion sequence of two-dimensional views in order to lead to an impression of 3D shape through structure from motion. The subjects studied the target while it was wob-

[1]This apparatus will be used throughout the remaining psychophysical experiments.

bling in a 5° spherical cone. In the subsequent testing phase, 4 objects were presented successively for testing: the target and the three corresponding distractors with x, y, z noise respectively. Across different trials, the order of presentation was randomized. After the presentation of each of the test objects, the subjects were required to make a target/distractor decision by pressing one of two appropriately-labeled keys. The test object remained on the screen for 3s or till the subject responded, whichever come first. A given target was never repeated for testing in order to reduce object dependent results. No feedback was given to the subjects as to their performance, and no practice session was used. Forty targets were tested in the experimental session that lasted about 30 minutes.

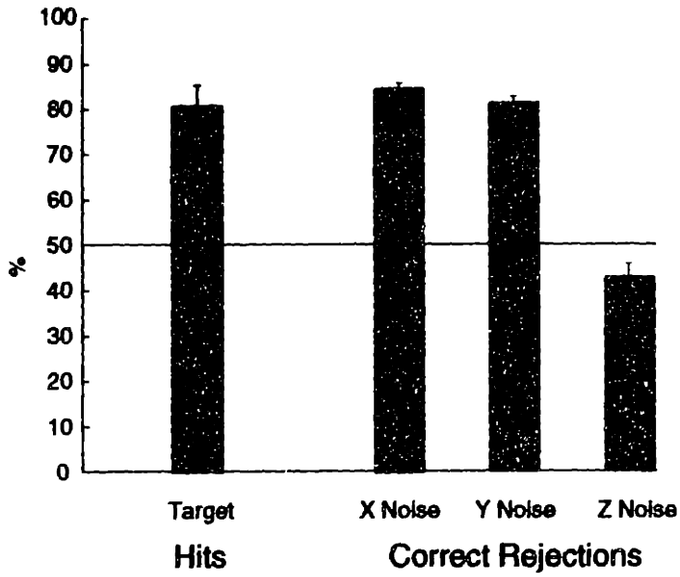

Figure 2-8: *Experiment 2:* Average percent correct is plotted in the left graph, while in the right graph we see how the performance changes with increasing noise.

Results

A single hit rate was computed for each subject, based on 40 same trials. Since the variable of interest was the direction of noise, the correct rejection rate was computed for each subject and for each level of noise direction (x, y, and z). Each correct rejection rate was based on 40 different trials. The mean hit and correct rejection rate averaged across 10 subjects are shown in Figure 2-8. The error bars indicate the

standard error of the mean computed across the 10 subjects.

A one-way Analysis of Variance (ANOVA) was conducted on the correct rejection rates with Noise Direction (x, y, z direction) as within subject variable.

Discussion

The threshold is not the same for all direction of noise. False alarm rates were higher for distractors where noise was added only to the z coordinates of vertices but significantly lower when the same amount of noise was added only to the x or y coordinates. This occured when the subjects are presented with maximal depth information. The representation seems to be view-based. Object representation is very sensitive to noise in the image plane, but much less to noise in the orthogonal direction. It could also be possible that the subjects were not used to this stereo presentation and the effect of noise in the z-direction was due to their misrepresentation of depth due to the particular set-up.

2.3.3 Experiment 3

We have seen that the representation for the direction orthogonal to the viewer is less precise even in the presence of all cues to depth. In the next experiment we intend to determine the relative importance of two cues: a more quantitative one like stereo an a more qualitative one like occlusion patterns.

Method

Subjects Fifteen paid subjects participated in this experiment (ten in group A and five in group B), all students at the Massachussetts Institute of Technology. All had normal or corrected-to-normal vision. All subjects were unaware of the purpose of the experiment.

Stimuli The stimuli were gray level images obtained as orthogonal projections of shaded 3D objects. A set of 10 wire-like objects was created from Class II defined in Appendix A. From each of these objects, another three objects (distractors) were

Figure 2-9: Example of objects used in Experiment 3.

generated, having the same orthogonal projections. In order to create these objects, we added noise at each vertex of the original object in the direction orthogonal to the projection plane. Thus, all the distractors have same 2D projection, but different 3D shape. All the objects were presented from a viewpoint causing a partial occlusion of one of the legs. This partial occlusion provided the subjects with qualitative depth information, that is ordinal depth order information. The original object and one of the distractors had the same depth ordering, while for the other two distractors, the noise was such that this depth ordering was reversed (see Figure 2-9). Two objects with the same occlusion pattern were not identical, the difference between the two was noticeable in 2D image by difference in the shading pattern. The difference was much more noticeable once the objects were presented in stereo. These 4 objects were considered as constituting an equivalence class.

Procedure We used the *match-to-sample* paradigm described in Section 2.2.2. In each trial, one equivalence class was tested. During the study phase, one of the four members of the class to be used as target for that specific trial was presented to the subjects undergoing a motion sequence for about 15s. This motion sequence of two-dimensional views was used so to lead to an impression of 3D shape through structure from motion. The subjects studied the target while it was wobbling in a 5^o spherical cone. In the subsequent testing phase, all the four objects constituting the class were presented successively. The test object remained on the screen for 3s, or till the subject responded, whichever come first. After presentation of each of the test

objects, the subjects were required to make a target/distractor decision by pressing one of two appropriately-labeled keys. Each subject completed a session composed of 40 trials, four for each of the equivalence classes. The test for the same equivalence classes were uniformly spread across the session in order to reduce object dependent results. Each of these trials used a different object in the class as target. There was no practice set and no feedback was given to the subject. Each session lasted about 30 min.

We subdivided subjects into two groups: group A viewed the stimuli binocularly, while group B received a stereoscopic presentation.

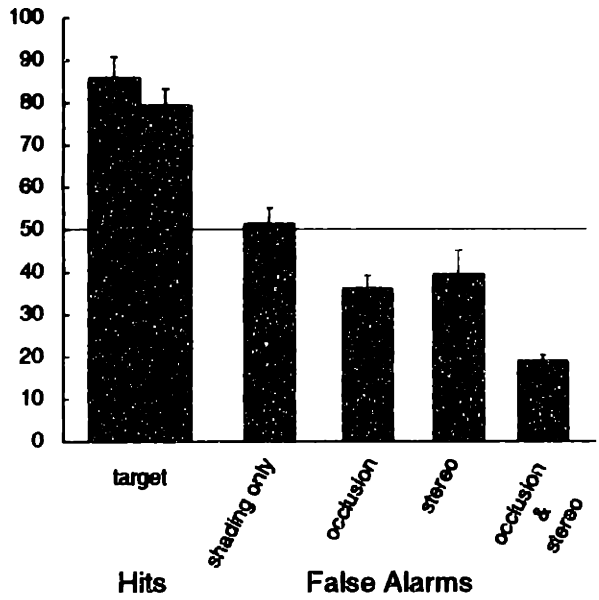

Figure 2-10: *Experiment 3:* Relative importance of qualitative and quantitative information. The horizontal line shows the chance level.

Results

The hit rate was computed for each subject based on 40 target trials. Two false alarm rates were computed for each subject, for each of the two levels of occlusion pattern: for the same occlusion, the False alarm rate was based on 40 distractor trials, while for different occlusion, it was based on 80 distractor trials. These values, averaged across 10 subjects for group A and 5 subjects for group B, are shown in Figure 2-10. The error bars indicate the standard error of the mean computed across the subject.

Discussion

Depth order information is used,but detailed information on depth dimension is not. The results could be influenced by the fact that distractors with opposite occlusion are often interpreted as if they were the target viewed from behind. The following experiment tests this hypothesis in more detail.

2.3.4 Experiment 4

Method

Subjects Six paid subjects participated in this experiment, all student at the Massachussetts Institute of Technology. All had normal or corrected-to-normal vision. All subjects were unaware of the purpose of the experiment.

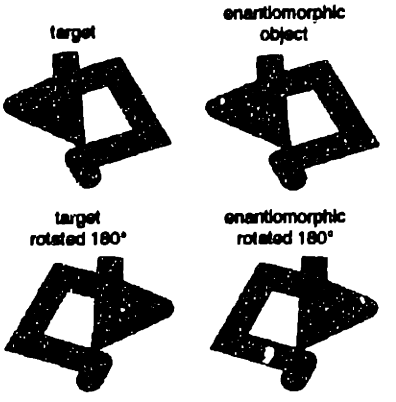

Figure 2-11: Example of objects used in Experiment 4.

Stimuli The stimuli were gray level images obtained as perspective projections of shaded 3D objects. A set of 20 wire-like objects to be used as targets was created from Class I defined in Appendix A. For each object, we constructed the mirror symmetric (enantiomorphic) correspondent to be used as a distractor. An example of the stimuli is shown in Figure 2-11.

Procedure We used the *match-to-sample* paradigm described in Section 2.2.2. During the study phase, the subjects were presented with the target undergoing a motion sequence for about 15s. This motion sequence of two-dimensional views was used in order to lead to an impression of 3D shape through structure from motion. The sub-

jects studied the target while it was wobbling in a 5° spherical cone. In the subsequent testing phase, four objects were presented successively for testing: the target and its enantiomorphic correspondent, both from a 0° view, and from the view obtained rotating the objects of 180° around their vertical axis. The order in which these four objects were tested was varied randomly from trial to trial. After each presentation of the test objects, the subjects were required to make a target/distractor decision by pressing one of two appropriately-labeled keys. The test object remained on the screen for 3s, or until the subject responded, whichever came first. For each object tested, in a subsequent trial the corresponding enantiomorphic object was tested also. These two trials were uniformly spread across the session. Thus, subjects run a complete session composed of 40 trials that lasted about 30 minutes. No practice set was used, and no feedback was given to the subjects on their performance.

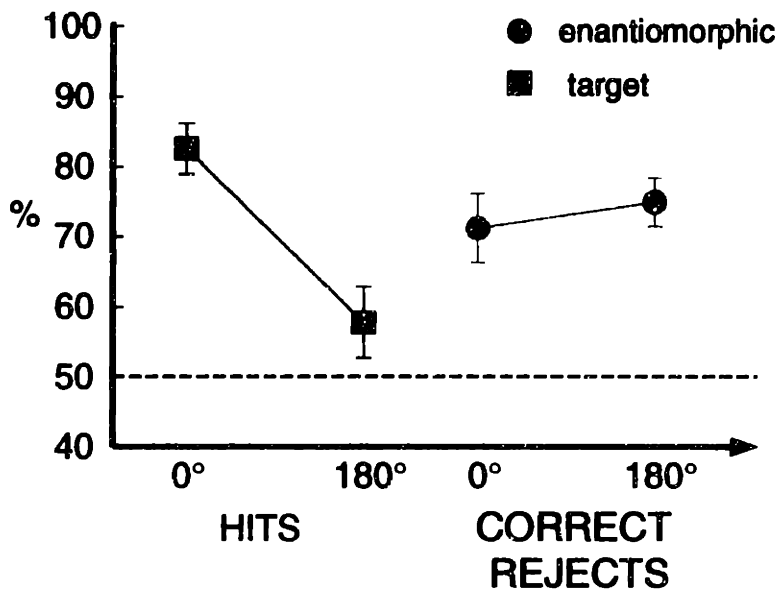

Figure 2-12: *Experiment 4:* Relative importance of qualitative and quantitative information.

Results

The hit rate and the correct rejects rate were computed for each subject at each of the two levels of rotation. All measurements were based on 40 trials, the former on target presentation, the latter on the presentations of the enantiomorphic objects. The mean hit and correct rejection rates are presented in Figure 2-12. The error bars

indicate the standard error of the mean computed across 10 subjects.

	Image	Occlusion
Target 0°	+	+
Target 180°	-	-
Mirror 0°	+	-
Mirror 180°	-	+

Table 2.1: Performance prediction from possible depth cues to recognition

Discussion

The subjects can reliably distinguish enantiomorphic objects from the unrotated version of the target. We notice that correct responses to target and distractors objects depend on two variable: image similarity and occlusion patterns (see Table 2.1. Only when both variables agree with the original presentation the performance is very good, while it decreases when one of the two variables is not in agreement. This suggest that the representation contains more than simple image information but codes some depth cues (like the occlusion).

2.3.5 Experiment 5

In this experiment we tested the time course of recognition with a set of distractors similar to those in Experiment 2.

Method

Subjects Sixteen paid subjects participated in this experiment. All had normal or corrected to-normal-vision. All subjects were unaware of the purpose of the experiment.

Stimuli The stimuli were gray level images obtained as perspective projections of shaded 3D objects. Three sets of 50 wire-like objects to be used as targets were

[1]This experiment could not be done without the help of Janine Mendola

created from Class II defined in Appendix A. Each set was used in one of three tests which varied the delay between the study and test phases of the experiment. Each of these targets was used to construct three comparison objects. For each object, noise was added to the x, y, z, coordinate of the vertices. The amount of noise varied from 4% to 40% of mean segment length, to create varying levels of difficulty. For each of the 50 trials, all of the three comparison objects were shown with their corresponding base object.

Procedure

Subjects' recognition memory was assessed with a four-alternative match-to-sample procedure. Subjects sat 109*cm* in front of the computer screen in a dimly-lit room. They were told that they would be presented with 5 wire figures. One figure appeared in the center of the screen, and four choices appeared in the corners of the monitor (which were numbered 1-4). Subjects' task was to pick verbally the corner object that best matched the sample object in the center of the screen. The test was administered 3 times with different stimuli. Initially, the sample and 4 choices were presented simultaneously. In the second version, the example stimulus appeared for 4 s and then disappeared. After a delay of 2 s, the 4 choices were presented. Finally, in a third version, the delay between presentation of the sample and choices was increased to 15s. During the delay, the subject and experimenter simply waited silently in front of the blank computer screen. There was no time limit to respond; subjects were instructed to make their best guess when unsure. They rested at several points during the test. There was no practice set. Subjects usually required 1 hour and 30 minutes to complete the entire procedure.

Results

The hit rate and three false alarm rates (corresponding to choosing one of the three possible distractors) was computed for each subject at each of three levels of delay for the monocular presentation and at each of two levels of delay for the stereo presentation. Each measure was based on 50 trials. The mean rates are presented in Figure 2-13. The error bars indicate the standard error of the mean computed across

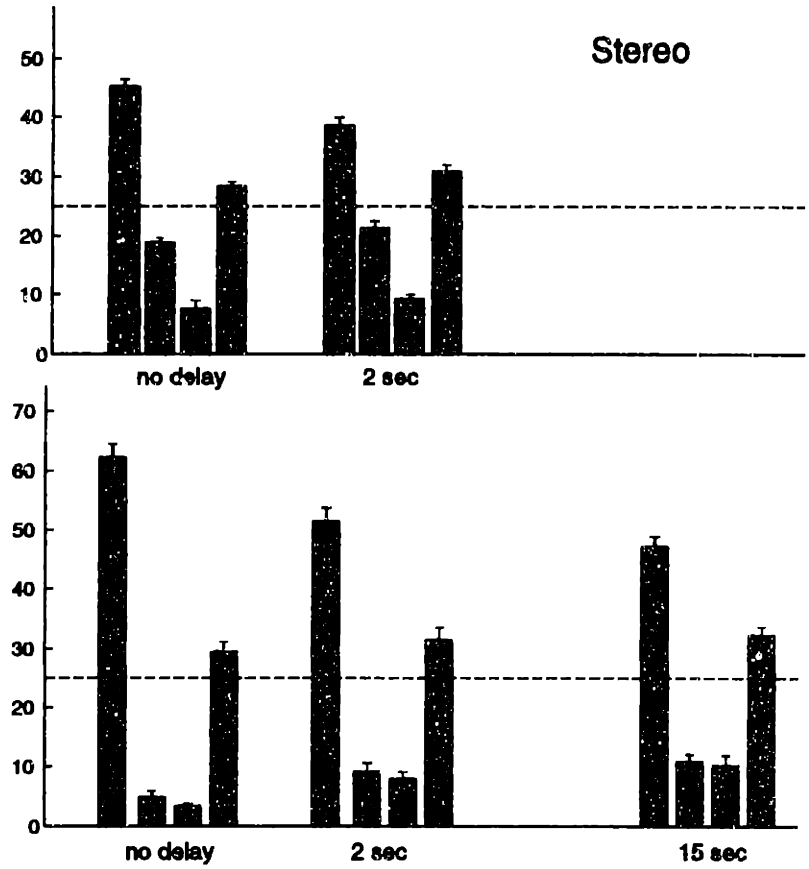

Figure 2-13: *Experiment 5:* Stimulus information decreases over time. Percent of correct recognition is shown in black, while percent of identification of the x, y, z noise distractor are shown in this order in gray.

16 subjects.

Discussion

Shape representation for recognition does not contain all the detail that could be possibly encoded. Recognition performance already decreases after only a couple of seconds. This rate of decrease is not maintained across, but slows down as it is shown by the behavior at 15 seconds.

2.4 General Discussion

Objects can be recognized by their projected shapes alone, even if these projected shapes vary greatly as the observed viewpoint changes. Current research is fo-

cused on whether this recognition is accomplished by matching an input shape to an orientation-dependent representation.

We addressed the issue of view-dependent representation, and whether this representation is systematically disrupted more by rotation in depth than by deforming the object image. In a number of experiment we tested recognition performance when objects have to be discriminated among other objects with varying degree of similarity. In Experiment 1, we tested recognition performance for a number of view of objects against distractors with varying degree of similarity. At low noise level, i.e. when objects are very similar, subjects are unable to discriminate, suggesting that the view tuning is obtained by storing a prototype that allows a certain amount of noise in the description. As the similarity decreases the recognition becomes easier but differentially so for different views of the object confirming that the frame of reference used in subordinate level classification is viewpoint dependent. In subsequent experiments (Experiment 2 to Experiment 5), we tested recognition of a single view of an object in more detail. The results suggest that, at least for the class of objects used in our experiments, this representation may be strongly configurational, weighting some characteristic parts of the image more than others.

Chapter 3

Object independent transformations

3.1 Effects of object retinal position

When we move about in the physical world, objects are projected on different positions on the retina, yet we recognize any object whether we see it in front of us or in the periphery of the visual field. The ability of the visual system to recognize patterns independent of where they appear on the retina is seldom questioned, but it should not be assumed. The insect visual system, for example, does not exhibit shift invariant properties. Flies store visual images at, or together with, fixed retinal positions, and can retrieve them only from there (Dill et al., 1993).

This observation brings us to question whether humans really *do* have a shift invariant representation for visual object recognition. Is it possible that this property is just a byproduct of the learning process? More specifically, do we need to have seen the object in one specific position before we are able to recognize it at that position, or can we extrapolate from one single example to anywhere in the visual field?

Unlike other transformations' effects, the effect of translation on recognition and identification of patterns or objects has seldom been studied, but the work that has been reported points toward a effect of translation of image patterns on recognition (Foster and Kahn, 1985; Nazir and O'Regan, 1990). In one of the first works on this

subject, Foster and Kahn showed that recognition performance drops linearly as the distance between two successively-presented patterns increases (Foster and Kahn, 1985). The stimuli consisted of simple random dot patterns composed of 10 dots randomly distributed within an imaginary 0.5° circle. Two stimuli were presented successively within a 1s interval, and subjects had to judge if the two were the same or different. The range of position varied from 1.0° left to 1.0° right of the fixation point. The effect seems very robust, and it is not likely to be attributable to other factors such as spatial attention or eye-movements. But in the analysis, the authors did not consider the ordering of stimulus presentation, confounding, for example, the case in which the first pattern was presented at the center of fixation and the second pattern in the left field, with the case in which the reverse was true. This creates problems when interpreting the results: it is difficult not to confound the influence of resolution with the effects of retinal shift.

This very nice linear effect has not been found with other type of stimuli, or with different paradigms. If the subjects undergo training before being tested (i.e they are taught to recognize the pattern at a given location), then the results are not straightforward. There is an effect of position most of the time, but it does not seem to be linearly dependent on the amount of displacement (Nazir and O'Regan, 1990). In this case, the stimuli were also different: dot stimuli resembling Chinese characters, where the dot size was much larger than in the previous experiment. A bigger range of displacements were tested, and the test was more controlled: the pattern was masked after being displayed, and an eye-tracker was employed to check eye movements. The results imply imperfect translation of the learned pattern to a new position.

As often seen in the recognition literature, we find a different result when implicit, rather than explicit, memory is tested. Biederman and Cooper have shown that degree of name priming is unaffected by a 4° shift (from 2° left to 2° right of the fixation point) (Biederman and Cooper, 1991). Experiments with contour deleted images reconfirmed these findings (Cooper et al., 1992a).

From these results, it seems likely that recognition is dependent on retinal position only in the case of difficult patterns. However these results are ambiguous, because

of the different tasks used in the above mentioned experiments. The invariance has been shown with tests on implicit memory, while the dependence is demonstrated with tests on explicit memory. In the following experiments, we examine whether, for objects belonging to the basic class described in Appendix A, shape representation is position-invariant in explicit memory.

3.1.1 Preliminary experiment

Translation seems to affect performance, and we have seen that the dependency of performance on position is highly dependent on the kind of patterns and the paradigm used. We are interested in understanding if the access to the representation of memcrized pattern is independent of the retinal position of the pattern when first learned. We will use a paradigm similar to the one used by Nazir and O'Regan (Nazir and O'Regan, 1990) with our chosen objects (see Appendix A)

Method

Subjects Seven paid subjects, all MIT students, participated in the experiment. All had normal, or corrected-to-normal vision. All subjects were unaware of the purpose of the experiment.

Apparatus The basic apparatus was similar to that used for Experiment 2. No stereo glasses were used. In addition, an IScan eye-tracker (RK-426 Pupil/Corneal Reflection Tracking System) was used in the phase of the experiment, when all the stimuli appeared at only one non-foveal position, so to check subjects' eye movements.

Stimuli A set of 50 objects were newly generated from the class described in Appendix A. Each stimulus was the orthographic projection of one of the 3D objects, and subtended about 0.9° of visual angle. The stimuli were rendered in black, and no lighting or shading information was available to the subject. Some of the stimuli used in this experiment are shown in Figure 3-1.

Procedure

The objects were tested in a two-alternative forced-choice (2AFC) paradigm. Short presentation (150 ms) time, together with masking, were used in both training

Figure 3-1: Stimuli used in the preliminary experiments: Three sample test stimuli and a mask are shown.

and testing phases to ensure eye movement preclusion. No shading information was used. Objects were viewed as two dimensional black patterns on a light background.

For each single trial, the fixation point appeared in the center of the screen for 500 ms to aid fixation, then the object was presented for 150 ms, followed by a 1 s mask presentation. The subject had to make an Yes/No decision by selecting one of two appropriately-labeled keys on the keyboard.

Learning was obtained through audio feedback: a beep informed the subject that an incorrect decision had been made. A score of 95% correct was required to end the training phase and start testing. Feedback about the overall performance was given to the subject every 60 trials. The testing procedure was identical to the training, but the audio feedback was eliminated.

Five presentation positions were utilized: at the center of the screen, or 2.4° left/right or up/down from that position. Only stimulus position was modified across trials, orientation and size were kept fixed. Each subject participated individually in one hour-long section. The section was subdived into ten blocks of trials. In each block, the subject learned to recognize one specific object (target) from all the other distractors during the training phase at *one* of the five possible positions. Audio feedback indicated to the subject that the previous decision had been incorrect. Once a level of 95% correct was achieved, in the subsequent *testing* phase, the subject was tested on recognition at all the possible positions. No feedback was given to the subjects in this phase of the experiment. There was no time limit on the response but subjects were asked to respond as quickly and as accurately as possible. Since

the subjects could not predict in which position the next test object was going to be presented (in contrast with the training phase), the use of eye-tracker was deemed superfluous.

The first ten members of the set were arbitrarily reserved to be used as targets. The rest of the objects in the set constituted the pool of distractors. The pool was large enough so that no distractor was ever shown twice in the presence of the same target. Target and distractors appeared with the same frequency (50% of the time each), and each position was tested equally often (six times per position) once learning was achieved. Of the ten different target objects that were tested for each session for each subject, two were first learned at each of the five positions. The order of learning was random. The sessions were identical for all subjects: they all learned the same objects at the same positions. Each session lasted approximatively an hour.

Results

Learning the discrimination was an easy task for the subjects, who took on average 66 training trials. The position of an object, whether study or test, was labeled as one of the following: center, down, left, right, up. Trials where reaction times were larger than 5 seconds were eliminated from the analysis. A sensitivity measure d_L was computed from the correct same trials (hits) and the incorrect different trials (false alarms) for each subject, for all possible combination of training and testing positions. Each value was based on 10 same trials and 10 different trials. Table 3.1 contains the values of d_L averaged across subjects.

	center	down	left	right	up
center	6.8	6.3	5.9	5.8	5.6
down	6.3	6.9	5.4	6.2	5.6
left	5.8	6.6	6.5	5.8	5.6
right	7.0	6.9	6.3	6.7	6.6
up	5.9	6.2	5.6	5.9	7.2

Table 3.1: Sensitivity d_L for all the possible combination of testing and training. Each line represents a different training position, each column a different testing position

The sensitivity measure d_L was submitted to a repeated measure analysis of variance treating testing and training size as within-subject factors. We found a main effect of both subjects ($F(6,144) = 10.1, p < 0.0001$), training ($F(4,144) = 2.8, p < 0.05$) and testing ($F(4,144) = 2.5, p < 0.05$) positions. There was no interaction effect between the last two variables ($F(16,144) = 1.6$).

Discussion

In the case of an effect of translation on sensitivity, we would have expected a strong interaction effect that is not seen in our analysis. There seems to be some effect of training position on the results. The result suggests that a translation-invariant representation is computed from a single image. However, there are still some problems with this paradigm. It ensures that the subjects have learned the target at a given position, but retesting each position more than once for each object renders the results ambigous. In the next experiment, we changed the paradigm so to try to eliminate these possible sources of confusion.

3.1.2 Experiment 6

Method

Subjects Ten paid subjects, all MIT students, participated in the experiment. All had normal or corrected-to-normal vision. All subjects were unaware of the purpose of the experiment.

Stimuli A new set of 180 stimuli was generated from the class described in Appendix A to be used as targets in the experiment. For half of these objects, one corresponding noisy distractor was generated. Added noise direction was picked randomly, and not limited to one of the cardinal axis. Each stimulus subtended at max 1.1°. The stimuli were rendered as orthographic projections of shaded 3D objects.

Procedure Fixation was aided by a computer-generated square black fixation point that subtended about 0.1°. Only three presentation positions were used in this experiment (center, 4.25° left and right of fixation). Stimuli were presented to subjects

in pairs in a sequential matching paradigm (see Section 2.2.1). At the beginning of each trial, the fixation point appeared in the center of the screen and stayed on for 500 ms to enable the subject to achieve fixation. After this time, one object appeared for $150ms$ either left, right, or on the fixation spot which remained on display for the whole length of the trial. After an interval of 1s, another object was briefly presented again for $150ms$ at one of the possible positions. The subject was required to judge whether the two objects were the same or different. Because of the brief presentation time, and the unpredictability of stimulus position, the eye-tracker was deemed superfluous. There was no time limit for the response, but the subject was asked to respond as quickly and as accurately as possible.

After the experimenter explained the paradigm, the subjects run a short session composed of 18 trials in order to practice with the task. Subsequently, each subject participated into three experimental sessions, one after the other. In a given run, no object was repeated in order to reduce object-dependent results. In the succeeding runs, in which the same objects were tested, the training position of each object was changed. Each run was composed of 180 trials, all position combinations occured 10 times each, both in same and different trial, so that a run consisted of 90 same and 90 different trials. Each session lasted about an hour, and no feedback was given to the subjects as to their performance.

	left	center	right
left	4.8	4.3	4.2
center	4.6	5.0	4.3
right	4.0	4.0	4.5

Table 3.2: Sensitivity d_L for all the possible combinations of testing and training. Each line represents a different training position, each columns a different testing position

Results

Because discrimination was determined by responses to both "same" and "different" objects, the discrimination index d_L was used to represent performance (Snodgrass

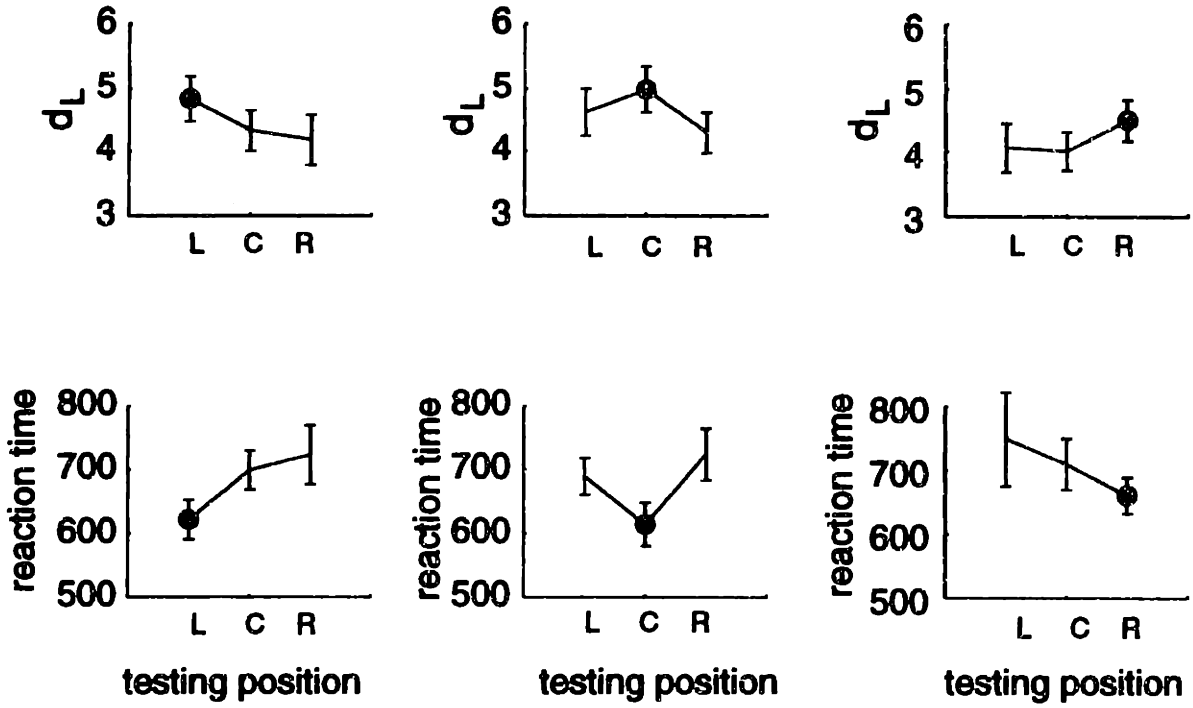

Figure 3-2: *Experiment 6:* Translation dependent recognition performance

and Corwin, 1988). The position of an object, whether study or test, was labeled as one of the following: center, left and right. Trials where reaction times were larger than 5*s* were eliminated from the analysis. A sensitivity measure d_L was computed from the correct same trials (hits), and the incorrect different trials (false alarms) for each subject, for all possible combinations of training and testing positions. Table 3.2 contains the values of d_L averaged across 10 subjects. The same results are presented in graphical form in Figure 3-2a. Error bars indicate the standard error of the mean computed across the 10 subjects.

Reaction times for same and different trials were also computed for all subjects, for all combinations of training and testing positions. Mean reaction times are presented in Figure 3-2b. Error bars indicate the standard error of the mean computed across the 10 subjects.

The sensitivity measure d_L was submitted to a repeated measure analysis of variance treating subjects, testing and training size as factors. We found again a main effect of subjects ($F(9,252) = 22.2, p < 0.0001$) and of training position ($F(2,252) = 3.5, p < 0.05$). There was no effect of testing position ($F(2,252) < 1$) and there was

an interaction effect between the two positions ($F(4, 252) = 3.3, p < 0.05$).

Discussion

The result suggests that a translation-invariant representation is computed from a single image. The difference between our outcomes and the results presented in the literature could be explained if the images were represented by encoding translation-invariant features (i.e., angles or segment length of wire-like objects), since these features are easier to extract from our connected objects than from carefully-chosen random-dot patterns.

3.1.3 Discussion

Our results show that the representation used for our objects is translation-invariant, and the results are apparently in contraposition with the results seen in the literature. The system appears to be unable to compute invariance to translation for the simple dot patterns, while it has no difficulty in the case of tubular objects. But the fact that the time required to learn the discrimination is so different in the two cases, seems to suggest that the dot-like patterns are intrinsically much more difficult to code. Some of the subjects were completely unable to learn the discrimination task, even when the pattern was presented foveally. The bent-wire objects, on the other hand, were easily coded and committed to memory, and this facility was reflected in the perfect transfer of information between the different regions.

It seems that the visual system is able to associate similar patterns of activation in different region of the visual field. The subjects must have committed a representation of the target to memory, and this representation is independent of the presentation position. Only where this pattern of activation is ambiguous is the transfer difficult and incomplete.

3.2 Effect of stimulus size on recognition

The extent of the projection of an object onto the retina depends on the size of the object itself and on its distance from the viewer. Recognizing familiar objects irrespective of their retinal projection is a task that is solved routinely by the visual system. This size-independent performance does not necessarily imply size-independent object representation. As in the case of shape, each object could be represented at a number of different scales and the combination of these different representations could give rise to the independent performance.

Despite the constantly varying retinal image, our visual world is characterized by constancy. The visual angle subtended by a given object varies greatly with the distance of the object from the viewer, but the object's perceived size remains unchanged. But does this size independent perception derive from a size independent object representation, or is each object represented at a number of different scales and the independent performance obtained by interpolation among these different representations?

Controlled psychophysical experiments have helped to get an insight on the representation of object size. At first sight, there appears to be no agreement on whether object representation is stored in a size-specific manner. All these experiments have used two-dimensional objects or line drawings as stimuli. In these cases, most of the results have shown a dependency of recognition performance on size changes.

When subjects were asked to determine whether two figures presented simultaneously were identical except for a change in size, reaction time measures for correct reaction show an increase as function of the size ratio between figures (Bundsen and Larsen, 1975). This result has been interpreted as if subjects, unable to discard size information, when asked to compare two objects of different sizes, retain one of the images as a reference and mentally resize the other image so that the two come in agreement. Each object is therefore stored with its size information, and the representation is not standardized before being committed to memory. These results have been related to Shepard's mental rotation results (Shepard and Metzler, 1971), but

size must be processed before identity of objects is obtained, because both test and control pairs show the same dependence on scale factor (Bundsen and Larsen, 1975).

Besner and his collaborators challenge this view, showing that different distractor types can produce a different slope for test and control pairs, therefore discarding the hypothesis of size transformation as preprocessing (Besner and Coltheart, 1975; Besner and Coltheart, 1976; Besner, 1983). They propose the existence of two parallel self-terminating systems, one that processes the "same" responses, and one that processes "different" responses. The first system is composed of a normalization stage that precedes the comparison stage, while the second is a feature analytic system (Besner, 1983).

The normalization could also be obtained in different ways. The coexistence of two different normalization stages has been proposed (Larsen and Bundesen, 1978). In the first case, the image is normalized to the perceptual reference system, while in the second case, the perceptual scale is altered so to fit the stimulus size. These two systems (discriminable by their time courses) subtend two different kind of memories: the short and the long term memory system respectively. Both memory systems could come into play in special conditions (Larsen and Bundesen, 1978).

In these early experiments, the images of the two objects to be compared were always simultaneously displayed. The subjects were viewing freely, and were able to modify or reinforce their object representation, by continuously shifting attention from one object to the other. They were not really tapping their memory representation. Later experiments also found the same dependency of reaction times on size of the object both when testing long term memory (Jolicoeur, 1987) and short term memory (Howard and Kerst, 1978; Werkhoven and Koenderink, 1993).

Mental dilation seems to be slower than mental constriction, suggesting that more processing time is needed to increase the "mental" size of a picture that to decrease it (Howard and Kerst, 1978).

The attempted normalization seems to be incorrect even for very simple features such as angles and speed of rotation (Werkhoven and Koenderink, 1993).

Contrary to what seems to be the case in the neuronal response, psychophysically-

perceived size, not retinal size, seems to affect perception (Milliken and Jolicoeur, 1992).

In testing size, it is important to use isolated objects. It has been shown that when drawing from memory, the presence of a background in the image reliably decreases the remembered size of an object present in a photograph (Legault and Standing, 1992).

In the literature of object size, there is additional evidence for the presence of two separate memory systems: one that is unaffected by metric changes and subserves implicit memory, and a second which is affected by metric changes and subserves explicit memory functions. (Cooper et al., 1992b; Biederman and Cooper, 1992; Cooper et al., 1992a; Fiser and Biederman, 1995). It has been suggested that these two systems reflect the separation of the "what" and "where" pathway described in the monkey (Mishkin et al., 1983; Biederman and Cooper, 1992). The proposal that the metrically-dependent representation is stored in the "where" pathway is not confirmed by other results. The idea that three-dimensional metrically-invariant structural descriptions subserving implicit memory system are stored in inferior temporal cortex seems to have stronger support (?). An alternative explanation excludes the possibility that the two systems compute in parallel, but suggests that the explicit memory system uses information from the structural description.

But is this normalization always taking place? Experiments in which subjects had simultaneously to rotate *and* scale images mentally, show no evidence of mental size scaling (Kubovy and Podgorny, 1981).

There is no neuropsychological evidence of specific brain lesions that selectively disrupt size or orientation perception. The only reported cases of size misperception are those of hemimicropsia following focal brain lesions (Cohen et al., 1994).

Neurons in the superior temporal sulcus selective for faces seem to have responses that are relatively invariant with respect to the size of the stimulus. Neuronal response is related to the retinal angle subtended by the stimulus. Few neurons were found to exhibit size constancy (Rolls and Baylis, 1986). In anesthetized monkeys, IT cells were found to be independent of size and position (Ito et al., 1995; Logothetis and

Pauls, 1995).

Therefore the question of size invariant is still debated. In the following section we will analyze whether in the case of wire-like objects the representation is size-dependent.

3.2.1 Experiment 7

Method

Subjects. Sixteen subjects participated in the experiment. They were students from the Massachussetts Institute of Technology who were paid for their participation. All subjects had normal or corrected-to-normal vision.

Stimuli. Two sets of 100 wire-like objects were created as members of Class II described in Appendix A. For each object in the second set, a corresponding distractor was generated, adding a variable amount of noise (in average 12% of mean segment length) to each target vertex. A mask was used: it was composed of 20 segments without any requirements on angle or intersection constrains. Examples of few objects and the mask is shown in Figure A-1.

Procedure. Each subject participated in one section lasting approximately 20 minutes. Before starting the session, the subjects were informed of the task by the experimenter. A *match-to-sample* paradigm was used, and each subject ran a session which consisted of 30 trials. The study phase started when the subject pressed a keyboard key, initiating the presentation of a fixation point for 500ms. After the disappearance of the fixation point an object, a new one for each block, was presented for five times for 150ms, followed by the mask at 1s intervals. In the study phase, the subject was asked to study the target. A message flashed on the screen advising the subject of the beginning of the test session that the subject could start by pressing an appropriately-labeled key. The subject recognition memory was tested 10 times in each testing phase. Each object presentation was preceded by the fixation point for 500ms, then an object that could be either the target or a distractor appeared for 150ms followed by the mask. The subject were asked to discriminate, as quickly

and accurately as possible, whether or not the object was the one he/she had just studied. No time limit on the response was imposed. There was no practice set, and no feedback was given to the subjects as to their performance.

The shapes could be presented at one of five sizes that were obtained by scaling the original object by a scale factor S, which could take one of the following values 1, 1.5, 2, 3, 4. The average visual angle subtended by the shapes was respectively 0.7, 1.0, 1.4, 2.0, 2.7 degrees. The target was always presented at the same size in all the blocks for all the subjects, and in this experiment, the intermediate size was used in the study phase, with scale factor 2. During the testing phase, the target and distractors could vary in size. Each target is tested only once for each of the 5 sizes. There were two sets of objects, and the subjects were randomly divided in two groups, each of which run the test on one of the two sets.

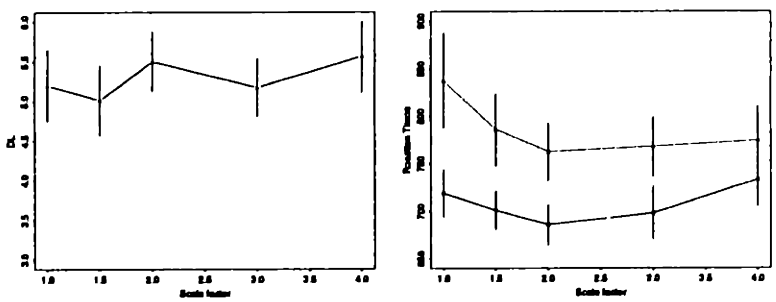

Figure 3-3: *Experiment 7:* Performance averaged across subjects is shown. (a) shows sensitivity d_L. Performance in the case where the size of the study and of the test object was identical is highlighted in grey. (b) shows reaction time data (means of medians) for correct responses for targets (squares and continuous line) and distractors (triangles and dashed line).

Results

The size of the test object was defined as the value S, by which the original object was scaled. The subjects' error rates were very small. We measured hit and false alarm rates across subjects, and we computed the sensitivity measure d_L from these values. For correct reactions to test objects, we computed the median reaction time for each subject, and for each value of size ratio, and then averaged the result across subjects. The results of this analysis are shown in Figure 3-3.

We found no effect of size ratio on sensitivity ($F(4,60) > 1$).

Discussion

No pattern of size dependence was found with these data. This independence could be explained by a ceiling effect. A possible solution is to use more difficult distractors. Another possibility is that subjects could learn invariances, since they were presented with more than one target size during testing. The next experiment was designed to address both these issues.

3.2.2 Experiment 8

Experiment 8 was designed to test effect of changes in object size on the recognition of novel three-dimensional objects, and of their corresponding two-dimensional projections.

Method

Subjects Forty-two MIT undergraduate students participated as volunteers in this experiment (15 in Experiment 8a, and 15 in Experiment 8b and 12 in Experiment 8c). Subjects were paid for their participation. All participants had normal or corrected-to-normal vision.

Stimuli The stimuli were grayscale images obtained as perspective projections of shaded 3D objects. A set of 180 wire-like objects to be used as targets was created from Class II defined in Appendix A. For half of these objects, we created a corresponding distractor, again adding noise to each of the vertexes (mean zero and standard deviation 12% of the average segment length) (see Figure 3-4).

Procedure The subjects' task was to judge whether two consecutively presented images represented identical objects. Each subject participated in one session, lasting approximatively 40 minutes, which consisted of four sessions. The first session was a practice, composed of only 10 trials, while the following three sessions were composed of 180 randomly-ordered trials. A *sequential matching* paradigm was utilized (see

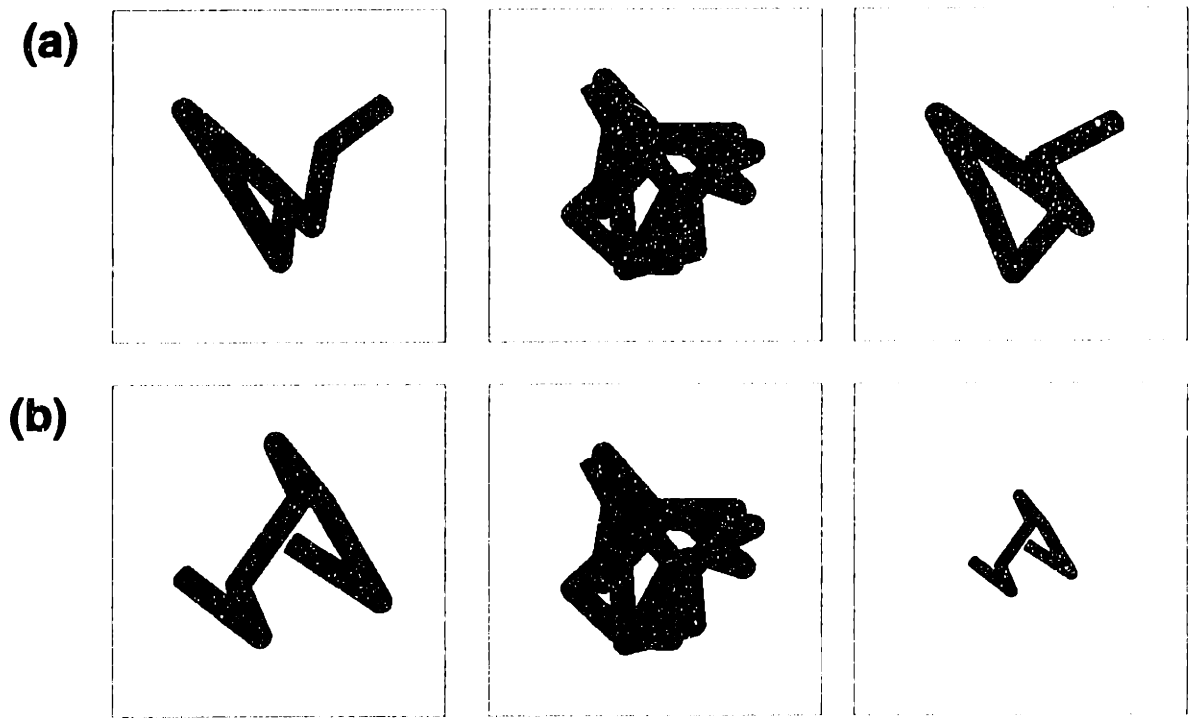

Figure 3-4: *Experiment 8:* (a) Example of different trial when study and test object have the same size. In this case the two objects have different shapes. (b) Example of same trial when the test object is half the size of the study object and has identical shape.

Section 2.2.1). Each trial started with the presentation of a fixation cross for 500 ms followed by the presentation of a study object for 6s, and then of a mask, which was displayed for 150 ms. One second after the mask disappeared, a second object was displayed and remained on the screen until the subject responded. The second object could be identical to, or different from, the first object. The subjects responded by pressing one of two labeled keys, choosing whether these two objects had identical shapes or not. No time limit was imposed on response and no feedback was given to the subjects as to the correctness of their response. Both objects could appear at one of three different sizes, obtained by scaling the original object by a factor S that could assume the values 1, 2 and 4. From the subject's viewing position, the objects subtended in average a 1.1° x 1.1° region in visual angle when presented at the smallest size.

The same 180 study objects were used in each experimental session, half of which were used in same trials (set S) and half were used in different trials with the corre-

sponding distractor (set D). These objects were divided in 3 subsets (a, b, c), each containing 30 objects from the S set and 30 from the D set. All the objects in a given subset were presented at one of the three scales in the study phase. The three subsets were rotated through all experimental conditions, resulting in a completely counterbalanced design during which each subset appeared equally often at a given scale in each cell of the main design. Every subject was tested ten times for each combination of study and test size, and trial type (same or different).

Three distinct experiments were run. In Experiment 8a, each object was displayed as a shaded three-dimensional object. When the size of a given object was modified, the whole object was scaled: in this condition, the thickness of the segments varied as the size varied. In Experiment 8b, the objects were still presented as shaded three-dimensional objects but the segments' thickness was maintained the same at all scales. Experiment 8c was identical to Experiment 8b, but the object were shown as black, eliminating all three-dimensional information available to the subjects. Each subject participated in only one of the three experiments.

Results

The size of an object, whether study or test, was defined as the value S (1,2 or 4) by which the original object was scaled. Trials where reaction times were larger than $5s$ were eliminated from the analysis. A sensitivity measure d_L was computed from the correct same trials (hits) and the incorrect different trials for each subject at each study and test size. For correct reactions, we also computed the mean reaction time for each subject at each study and test size for both same and different trials. The mean of these values across subjects, together with the standard error, are shown in Figure 3-5 for Experiment 8a and 3-6 for Experiment 8b and 3-7 for Experiment 8c.

Reaction times and sensitivity d_L were submitted to a repeated measure analysis of variance, treating size and test as a within subjects factor. In Experiment 8a, we found no effect of training or testing size on d_L ($F(2,112) < 1$), but a strong interaction effect between these two variables ($F(4,112) = 4.2, p < 0.005$).

In Experiment 8b, we found no effect of testing size on d_L ($F(2,112) < 1$), a weak

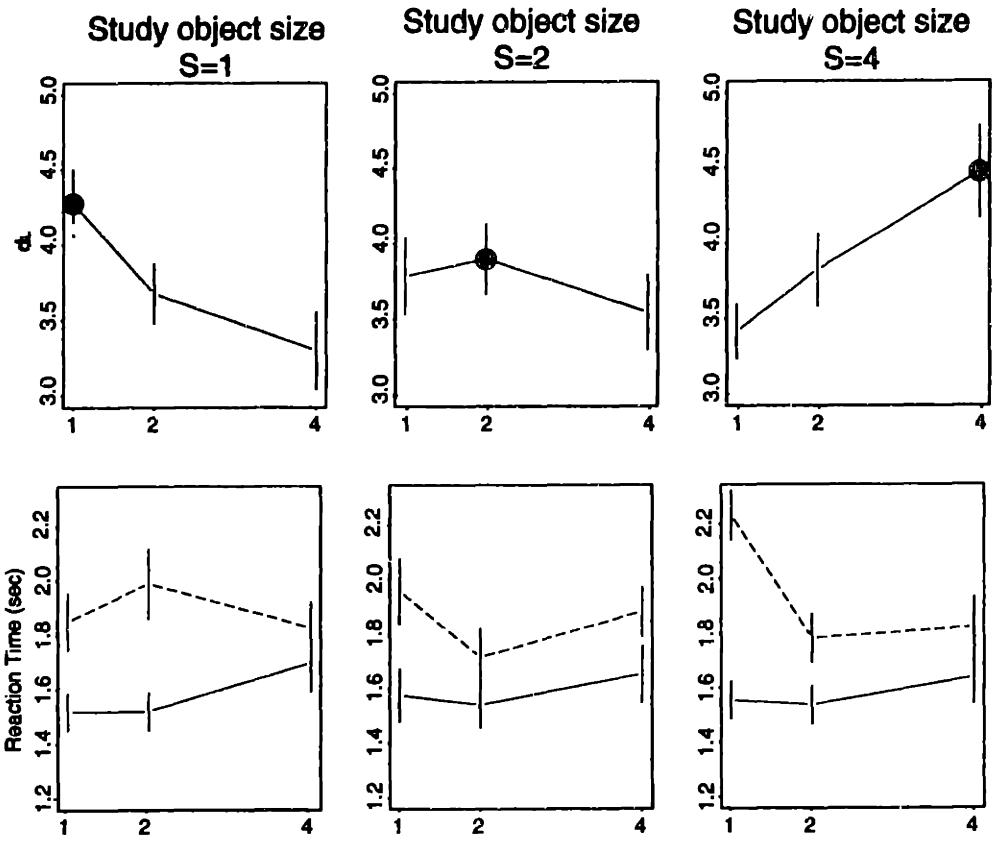

Figure 3-5: *Experiment 8a:* (a) Sensitivity, d_L plotted as function of test object scale, one plot for each different study object size. (b) Reaction times for correct trials (same: continuous line, different: dashed line). One plot for each different scale size at study.

effect of training ($F(2,112) = 3.6, p < 0.05$) and a strong interaction effect between these two variables ($F(4,112) = 4.5, p < 0.005$).

The pattern of results is seen in Experiment 8c is similar to Experiment 8a; there is still an interaction effect on the sensitivity d_L ($F(4,64) = 8.6, p < 0.0001$) but no other main effect.

Discussion

Objects shown at the same size in both study and test phase were recognized more quickly, and more accurately, than those that were shown at different sizes. It is interesting to note that this effect was shown also in the different trials, as would be expected, with distractors very similar to the targets.

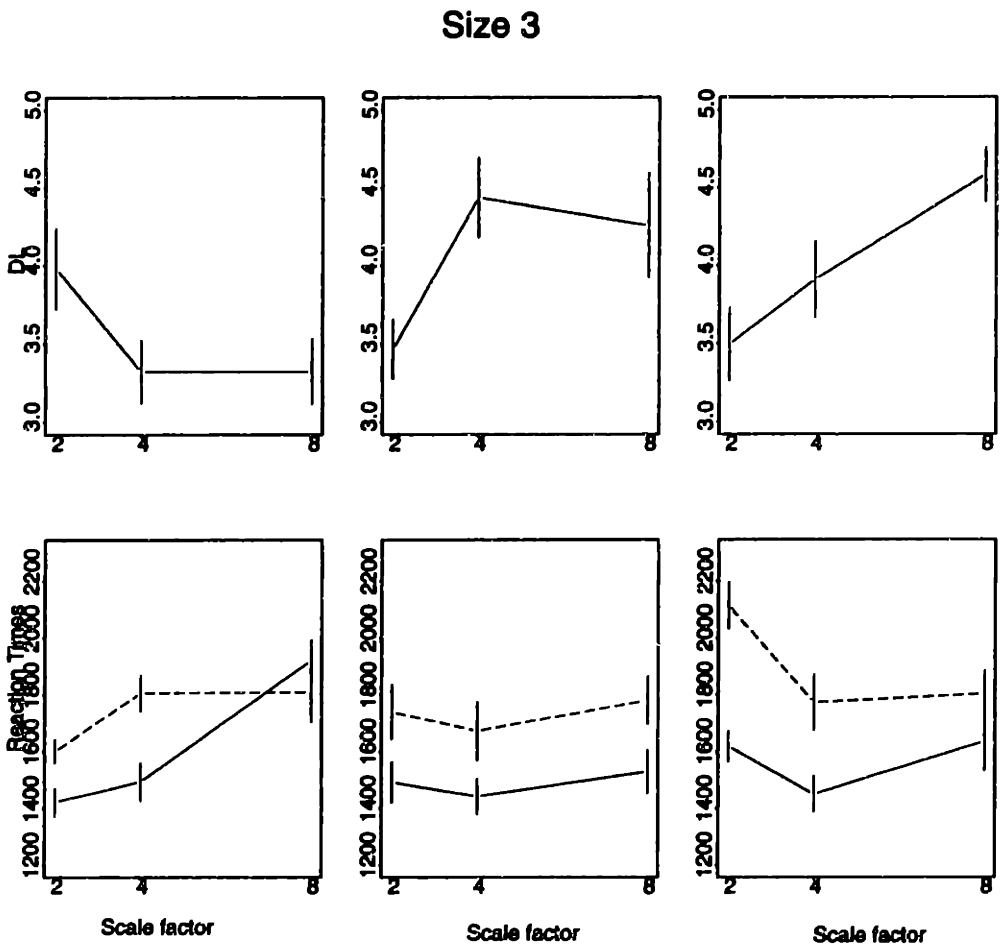

Figure 3-6: *Experiment 8b:* We show here performance averaged across subjects. The first row of plots shows Sensitivity, d_L plotted as function of test object scale, while the second shows mean and standard deviations of hit rate (continuous line) and false alarm rate (dashed line) for all tested sizes. The third row shows reaction time data (means of medians) for correct responses for targets (continuous line) and distractors (dashed line).

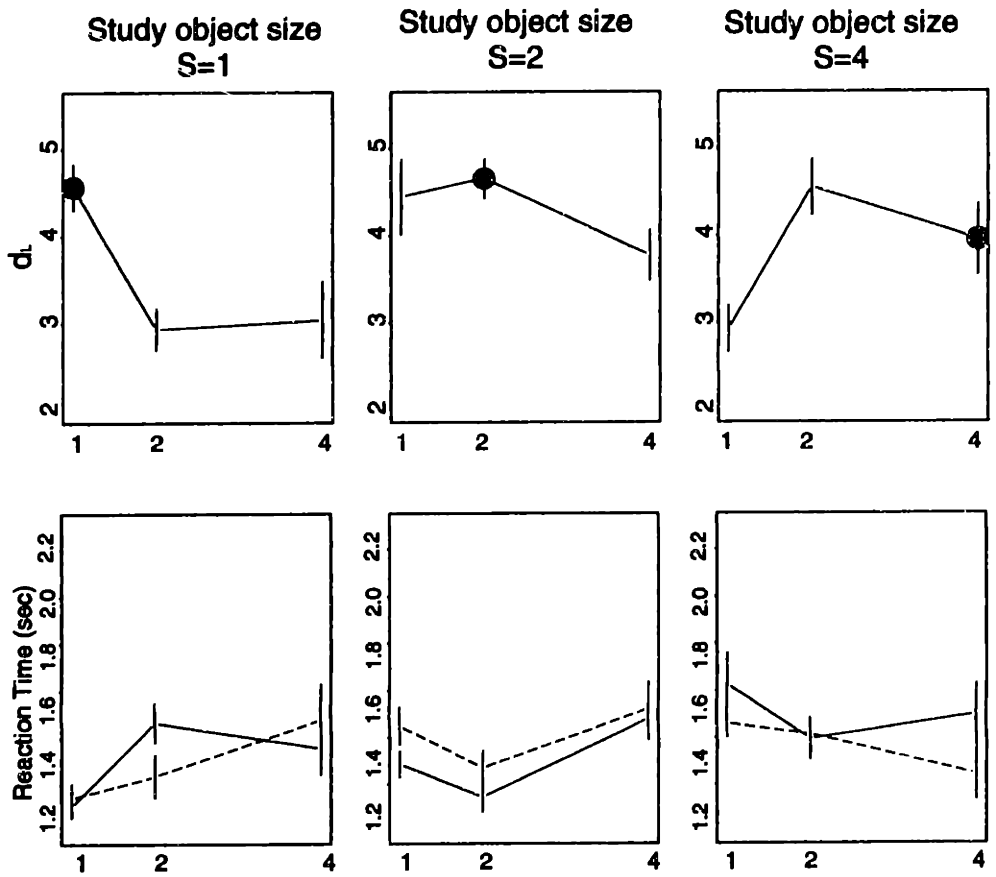

Figure 3-7: *Experiment 8c:* (a) Sensitivity, d_L plotted as function of test object scale, one plot for each different study object size. (b) Reaction times for correct trials (same: continuous line, different: dashed line). One plot for each different size used in the study phase.

3.2.3 Experiment 9

Experiment 9 was designed to test effect of changes in both object size and object rotation on the recognition of novel three-dimensional objects.

Methods

Subjects. Fifteen MIT undergraduate students participated as volunteers in this experiment. Subjects were paid for their participation. All participants had normal or corrected-to-normal vision.

Stimuli. A set of 32 objects was created from the Class II described in Appendix A. For half of these objects, we created a corresponding distractor, again adding noise to each of the vertices (mean zero and standard deviation 12% of the average segment length) (see Figure 3-4).

Procedure Each session consisted of 32 trials in a *sequential matching* paradigm (see Section 2.2.1). In the study phase, the fixation point appears for 500ms, then the target wire is presented in a rotation sequence that lasts approximatively 15*s*. In the test phase, the fixation point appears for 500ms, then a wire that could be either the target or one of many distractors appears for a maximum of 3*s* (or less, if the subject answers before then). There are ten tests during the test phase. Subjects have to press one of two keys to say whether the object presented is the current target or not. There was no time limit within which to respond. Each object could be presented at a number of scales and degrees of rotation. When presented at the the smallest scale, it was subtending about 1 degree of visual angle. It can be scaled by the following factors: 1, 1.5, 2 (used as training size), 3 and 4. The object could also be rotated around its horizontal or vertical axis of -45, -25, 0 , 25, 45 degrees. During this second phase, the target and distractor can vary in both size and rotation. Each target is shown in the testing phase only once for each of the 5 sizes, and once for each of the 9 different rotations. The experimental design insured that all the possible combinations are tested equally often. There were ten repetitions for each possible condition. No practice set was used, but the first trial was discarded from the analysis.

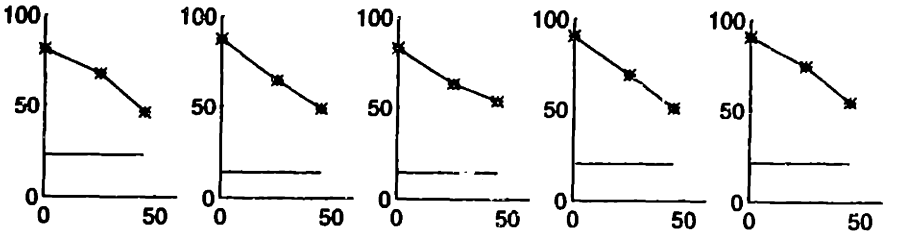

Figure 3-8: *Experiment 9:* View dependency and size: Hit rate and false alarm as function of size are plotted separately for different viewpoints. The target with size one is about 1 degree of visual angle. It can be scaled by the following factors: 1, 1.5, 2 (used as training size), 3 and 4. The test object can be rotated of $\pm 25, 45^{o}$ around vertical or horizontal axis.

Results

The first and last block were eliminated from the analysis. The size of the test object was defined as the value S, by which the original object was scaled, and its rotation by the angle A, of which it was rotated around any axis. There was no effect of axis of rotation, so we collapsed all the data in three bins given by the absolute value of the rotation. We measured hit and false alarm rates across subjects, and from these, we computed the sensitivity d_L. For correct reactions to test objects, we computed the mean reaction time for each subject, for value of size ratio, and for each value of angle of rotation, and then averaged across subjects. The results of this analysis are shown in Figure 3-8. There is a main effect of rotation ($F(2, 126) = 47.6, p < 0.001$) but no effect of size or any interaction among the two.

Discussion

There are two explanations for why the size effect does not appear in this experiment. The most appealing argument would suggest that size is processed independent of viewpoint and therefore its effect does not appear when testing recognition in this manner. But unfortunately we have also to take into account the fact that there is a possibility, as we have seen with Experiment 7, that we do not find an effect due to the testing procedure. To exclude this possibility new experiments are in progress.

Part II

Monkey Physiology

Chapter 4

Inferior temporal cortex and object recognition

In order to characterize the representation used by the brain to encode three-dimensional (3D) objects, one of the first issues to address is whether the frame of reference used is object-centered or viewer-centered. In the preceding chapters, we reported evidence which supports the idea that objects are represented in the brain as a collection of two-dimensional (2D) views.

In this chapter, we provide evidence for a similar view-dependency of recognition in the non-human primate. In combined psychophysical and electrophysiological experiments, we tested monkeys for the ability to generalize recognition for views generated by rotating objects around any arbitrary axis. Throughout this process, the activity of single units in inferotemporal (IT) cortex was recorded. Recognition at the subordinate level becomes increasingly difficult for the monkeys, as the stimulus is rotated away from a familiar attitude. Neural recordings revealed a small population of neurons (6% of the recorded ones) with remarkable selectivity for individual views of objects which the monkeys had learned to recognize. Plotting the spiking rates of neurons as a function of rotation angle results in systematic view-tuning curves for rotations in depth. For some of the tested objects, we found different neurons tuned

[1]Portions of this chapter have appeared before in Bricolo et al. (1995). This research was done in collaboration with J.Pauls and N.K. Logothetis at Baylor College of Medicine, Houston, Texas.

to different views of the same object.

These results are in agreement with a recognition model that accomplishes view-invariant performance by storing a limited number of object views or templates together with the capacity to interpolate between the templates (Poggio and Edelman, 1990).

4.1 Theories of Object Recognition

Most theories of object recognition assume that the visual system stores a representation of an object, and that recognition occurs when this stored representation is matched to its corresponding sensory representation, generated from the viewed object (Ullman, 1989). This assumption, however, raises two obvious questions. Are objects represented explicitly in the visual cortex, say, by the activation of a set of selective neurons on the top of a visual processing hierarchy, or are they implicitly represented by the activity of large populations of cells, each of which might have little selectivity for any of the complex features of an object? Furthermore, are the stored representations object-centered, 3D descriptions of the objects, or are they viewer-centered descriptions, corresponding to two-dimensional perspective views?

Optimal object representations may vary for different recognition tasks. Objects are usually recognized first at a particular level of abstraction, called the *basic level.* For example, a *Golden retriever* is more likely to be first perceived as a *dog*, rather than as a *retriever* or a *mammal.* Classifications at the basic level carry the highest amount of information about a category, and are usually characterized by distinct shapes (Rosch et al., 1976). Classifications above the basic level, *superordinate categories*, are more general, while those below the basic level, *subordinate categories*, are more specific, sharing a great number of attributes with other subordinate categories, and having to a large extent similar shape (Rosch et al., 1976; Jolicoeur et al., 1984).

In primate vision, shape is considered to be the critical attribute for object recognition. Psychophysical studies have shown that the recognition of objects is typically unaffected in gray-scale photographs, line drawings, or in cartoons with wrong color

and texture information. Evidence as to the importance of shape for object perception comes also from clinical studies showing that the breakdown of recognition, resulting from circumscribed damage to the human cerebral cortex, is most marked at the subordinate level, at which level the greatest shape similarities occur (Damasio, 1990).

Models of recognition differ in the spatial frame used for shape representation. Current theories using object-centered representations assume either a complete 3D description of an object (Ullman, 1989), or else a structural description specifying the relationships among viewpoint-invariant volumetric primitives (Marr, 1982; Biederman, 1987). In contrast, viewer-centered representations model 3D objects as a set of $2D$ views, or aspects, and recognition consists of matching image features against the views in this set.

Since a recognition system based on $3D$ descriptions cannot easily be discerned from a viewer-centered system exposed to a sufficient number of object views, when tested against human behavior, both models predict well the view-independent recognition of familiar objects (Biederman, 1987). However, only viewer-centered representations, can account for performance in recognition of various kinds of novel objects at the subordinate level (Rock and DiVita, 1987; Tarr and Pinker, 1990; Edelman and Bülthoff, 1992). Although viewer-centered representations have been often considered implausible due to the vast amount of memory required to store all discriminable object views needed in order to achieve viewpoint invariance, recent theoretical work has shown that a simple network can achieve viewpoint invariance by interpolating between a small number of stored views (Poggio and Edelman, 1990). Computationally, this network uses a small set of sparse data corresponding to an object's training views, so as to synthesize an approximation to a multivariate function representing the object. This approximation technique is known by the name of Generalized Radial Basis Functions (GRBFs) (Poggio and Girosi, 1990b). A special case of such a network is that of Radial Basis Functions (RBFs). RBFs are conceived of as "hidden-layer" units, the activity of which is a radial function of the disparity between a novel view and a template stored in the unit's memory. Such an interpolation-based net-

work makes both psychophysical and physiological predictions which can be tested directly against behavioral performance and single cell activity.

4.2 Representation of Objects in Inferior Temporal Cortex

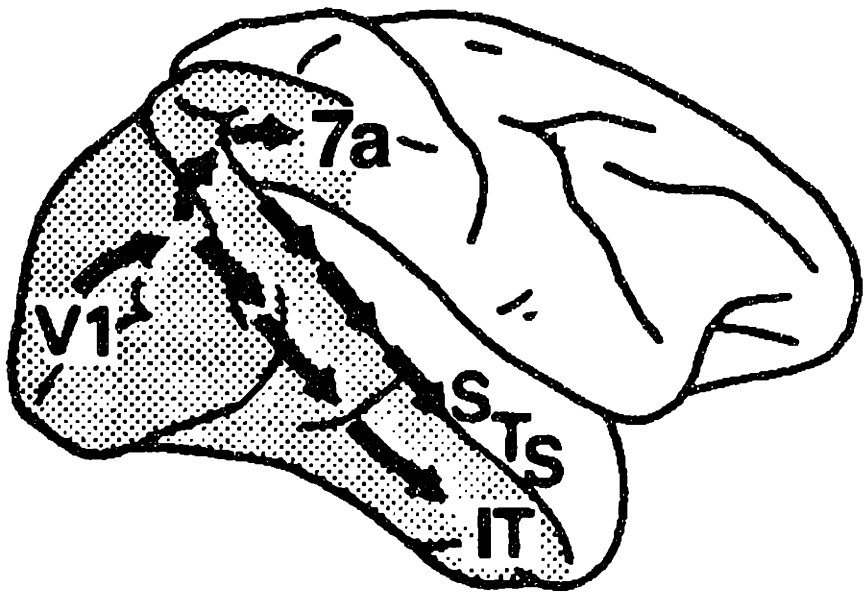

Figure 4-1: Lateral view of temporal lobe of a macaque monkey. V1: primary visual cortex, STS: superior temporal sulcus, IT: inferior temporal cortex (Adapted from Ungerleider and Mishkin, 1982)

After reaching area V1 (the primary visual cortex), the signal coming from the retinas is analyzed separately through two different functional pathways (Ungerleider and Mishkin, 1982). The *where* pathway proceeds dorsally into the posterior parietal area and seems to be concerned with location and spatial relationship amongst objects. The *what* pathway —the one we are concerned with— is believed to process form. The signal proceeds from area V1 ventrally through V4 and IT cortex, the highest area in the hierarchy that responds only to visual stimuli (see Figure 4-1). IT cortex in monkeys is therefore one possible brain area in which to explore the existence of cells selective for views of novel objects.

IT cortex corresponds to the areas 20 and 21 of Brodmann. It extends from a

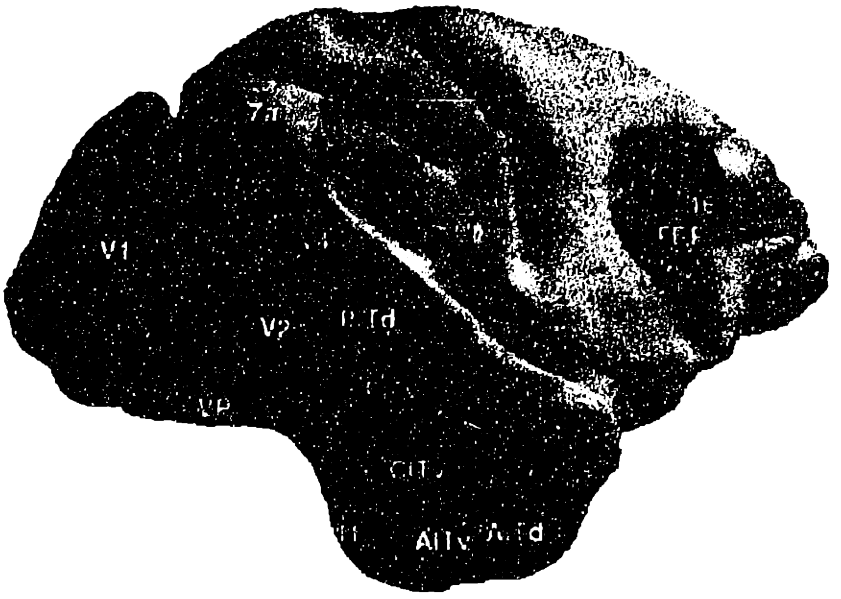

Figure 4-2: Lateral view of temporal lobe of a macaque monkey. In blue shades are the areas composing the *what* pathway. (Adapted from Carman et al., 1995)

couple millimeters anterior to the inferior occipital sulcus to just posterior to the temporal pole, and from the fundus of the superior temporal sulcus (STS) to the fundus of the occipito-temporal sulcus (Figure 4-1). Area IT is roughly coextensive with area TE as described by von Bonin and Bailey (1947), which was later subdivided into areas TE anteriorly and TEO posteriorly (Von Bonin and Bailey, 1947).

IT cortex has been shown to be essential for object vision. Patients undergoing unilateral anterior temporal lobectomy for the relief of focal epilepsy exhibit specific visuo-perceptual deficits (Milner, 1958; Milner, 1968; Milner, 1980; Kimura, 1963; Lansdell, 1968), and significant impairment in memory of complex visual patterns (Milner, 1980; Kimura, 1963; Milner, 1968; Taylor, 1969). Similarly, bilateral removal of IT cortex in monkeys yields marked impairment both in the retention of visual discrimination habits acquired before the surgery, and in the postoperative acquisition of new discriminations (Gross, 1973; Iwai, 1969; Cowey and Gross, 1970). Such lesions to area TEO, and to area TE, yield disruptions of pattern perception and recognition, while leaving thresholds for low-level visual tasks unaffected (Iwai, 1969; Gross, 1973).

Electrophysiological data first reported by Charles Gross and colleagues has provided further evidence regarding the role of IT cortex in object recognition (Gross

et al., 1967; Gross et al., 1969). The data showed that the posterior part of area IT has a rough visuotopy while the anterior part is not visuotopically organized. IT neurons have large, ipsilateral, contralateral, or bilateral receptive fields, that almost always include the fovea. Most neurons are selective for stimulus attributes, such as size, shape, color, orientation, or direction of movement (Gross et al., 1969; Gross et al., 1972). Some neurons respond best to complex shapes, including hands, trees, and human or monkey faces. A large number of investigations confirmed and extended these initial findings. As it stands, we know that IT neurons respond only to visual stimuli (Gross et al., 1967; Gross et al., 1972; Desimone et al., 1984). In general, neurons responsive to similar features seem to be organized in columns spanning most of the IT cortical layers (Tanaka et al., 1991; Fujita et al., 1992). Neurons recorded on the same electrode tend to have similar stimulus selectivity, and are more likely to show functional interactions than those recorded on different electrodes spaced farther apart (Gochin et al., 1991).

The activity of IT cells has been studied by various researchers using three different basic types of stimulation:

1. *Natural or hand-made common objects.* IT neurons do respond in a selective manner to the shape of various natural or man-made objects (Desimone et al., 1984), and maintain this selectivity even when the stimuli are defined by visual cues other than luminance or color contrast. For example, neurons in this area respond selectively to shapes defined through relative motion or texture differences (Sáry et al., 1993).

2. Sets of *complete orthonormal functions.*

    i) Fourier Descriptors. These parametric shape descriptors are typically used in computer pattern recognition systems for describing boundary curvature. IT neurons were found to be selective for an individual Fourier Descriptor, although most cells responded at different rates to a variety of other stimuli as well (Schwartz et al., 1983). However, when neurons were tested with patterns composed by more than one descriptors, their

response could not be predicted from the cell's responses to each member (Albright and Gross, 1990). This finding suggests that objects in IT are unlikely to be encoded through a set of general shape descriptors.

ii) Walsh functions. These are functions that can be used to synthesize any two-dimensional visual pattern to a required degree of accuracy (Richmond et al., 1987). Neurons were found to respond in different degrees to individual members of such sets.

3. *"Empirical" decomposition.* In these experiments, the stimuli are stripped down of most attributes to arrive to the "minimum" shape that still elicits a response. In general, when this technique is employed, we see that neurons responsive to similar features seem to be organized in columns that span most of the cortical layers (Tanaka et al., 1991; Gochin et al., 1991; Fujita et al., 1992). Neurons recorded on the same electrode tend to have similar stimulus selectivity, and are more likely to show functional interactions than those recorded on different electrodes spaced farther apart (Gochin et al., 1991).

The most striking class of highly selective cells in IT are those responding to the sight of faces (Bruce et al., 1981; Perrett et al., 1982; Cormack, 1985; Hasselmo et al., 1986; Yamane et al., 1987), or to parts thereof (Perrett et al., 1987; Young and Yamane, 1992b; Young and Yamane, 1992a). Face cells were found deep in the lower bank and fundus of the superior temporal sulcus (STS), and within the superior temporal sulcus in the poly-sensory area, located immediately dorsal to IT cortex in the fundus and in the upper bank of the sulcus (Desimone et al., 1984; Perrett et al., 1982). Since no other area in IT has such a high concentration of face-selective cells, this part of temporal cortex is thought to be specialized for the analysis of faces. Moreover, this part of the cortex is interconnected with the amygdala, where face-selective cells were also recorded (Sanghera et al., 1979). Most face-selective neurons are two to ten times more sensitive to faces than to other complex patterns, simple geometrical stimuli, or real $3D$ objects (Perrett et al., 1979; Perrett et al., 1982). Presenting different views, or parts of a face in isolation, revealed that the neurons

may respond selectively to face views (Desimone et al., 1984; Perrett et al., 1985), features, or subsets of features (Perrett et al., 1987; Young and Yamane, 1992b). Thus, face neurons do have properties reminiscent of an RBF network showing selectivity to specific views. Is such view-selectivity specific to faces? Could one expect to find neurons in area IT that are tuned to views (or parts thereof) of *nonsensical* objects, that the monkey learns to recognize?

The term "object constancy" is used to denote the intriguing ability of the biological visual systems to recognize objects under greatly different illumination and viewing conditions. The basis of such "object-related" and "condition-independent" perceptual performance might be the ability of neurons in IT cortex to maintain their selectivity over a variety of affine transformations.

The selectivity of IT neurons is indeed maintained over changes in stimulus size, contrast, color, or orientation in the frontal plane (Gross and Mishkin, 1977). Changes in stimulus size normally cause changes in the absolute firing level of the neuron, which indicates that the absolute response of an IT neuron rarely exhibits "size constancy" (Schwartz et al., 1983). The relative preference for a particular stimulus, however, is maintained over changes in size and position within the receptive field, and to this extent IT neurons can be thought of as showing size and position invariance.

Cell responses to different views generated by rotations in the frontal plane may be view-dependent (Tanaka et al., 1991) or independent (Desimone et al., 1984). Responses for rotations in depth of an object, on the other hand, have not been studied systematically. Face-selective cells have been reported where response was affected by the rotation of the head around the vertical axis (Desimone et al., 1984; Perrett et al., 1985). Reports also exists of cells responding independently of the orientation of a moving head in relation to the viewer (Hasselmo et al., 1986). No data are available regarding the selectivity of the neurons for rotation in depth of non-face objects, or alternatively regarding the ability of the monkeys to generalize recognition over different orientations for novel objects encountered from a limited number of viewpoints.

Rotation of the head around the vertical axis affected many cells. Cells were

found to be selective for a particular view of the head. In particular, some cells were maximally sensitive to the front view of a face, while their response rate fell off as the head was rotated into the profile view. Other cells were sensitive to the profile view, with no reaction to the front view of the face. A detailed investigation of these type of cells by Perrett et al. (1985) revealed five types of cell in the superior temporal sulcus, each cell type being maximally responsive to one view of the head. The five types of cell were separately tuned for full face, profile, back of the head, head up, and head down. In addition, two subtypes have been discovered, which respond only to left profile, or right profile, confirming that these cells are involved in visual analysis rather than representing specific behavioral or emotional responses. Most of these neurons were found to be two to ten times more sensitive to faces than to simple geometrical stimuli or $3D$ objects (Perrett et al., 1979; Perrett et al., 1982). Masking out, or presenting parts of the face in isolation, revealed that different cells responded to different features or subsets of features. Different faces, however, failed to elicit differentiated activity of the cells suggesting that this cell population was encoding the object "face", rather than specifying the presence of particular faces.

In related psychophysical work, Rosenfeld and van Hoesen provided some evidence of perceptual generalization to different views of faces by monkeys (Rosenfeld and Van Hoesen, 1979). In their study, monkeys were able to discriminate faces of other monkeys and made differential responses to them as readily as they did to simple geometric patterns. Furthermore, once an initial full face discrimination was made, alterations in stimuli entailing manipulations of posture, orientation, color, size, and illumination had very little effect on performance (Rosenfeld and Van Hoesen, 1979).

Recently, a quantitative study using correlation analysis between the quantified facial features and the neurons' responses has shown that face neurons can detect the combination of the distances between facial parts, such as eyes, mouth, eyebrows, hair, etc. (Yamane et al., 1988). These cells show a remarkable redundancy of coding characteristics, as it becomes evident by the fact that merely two dimensions were found to be enough to explain most of the variance in a population of studied neurons. For example, all the width measurements, such as the width of the eyes or the mouth,

the inter-ocular distance, etc., covary with the general width of the face. Moreover, the neurons responsive to faces exhibited graded responses with respect to the face stimuli, with each cell appearing to participate in the representation of many different faces (Young and Yamane, 1992a). Thus, face neurons do have properties reminiscent of an RBF network. Is such view-selectivity specific to faces? Could one expect to find neurons in this area that are tuned to views (or parts thereof) of *nonsensical* objects, that the monkey just learns to recognize?

In the experiments described below, we trained monkeys to recognize novel objects presented from one view, and subsequently tested their ability to generalize recognition for views generated by mathematically rotating the objects around arbitrary axes. We have then examined whether neurons in IT cortex respond selectively to novel objects that the monkey learns to recognize, and whether or not those cells that might be selective show view dependent activity.

4.3 Materials and Methods

4.3.1 Subjects and Surgical Procedures

Three juvenile rhesus monkeys (*Macaca mulatta*) weighing 7-9 kg were tested in combined psychophysical and physiological experiments. The animals were cared for in accordance with the National Institutes of Health Guide, and the guidelines of the Animal Protocol Review Committee of the Baylor College of Medicine.

Each monkey was first trained to sit in a primate chair. The animals underwent a surgery for the placement of a head restraint post, and a scleral-search eye coil (Judge et al., 1980) to measure eye movements. The monkeys were given antibiotics (Tribrissen 30 mg/kg) and analgesics (Tylenol 10 mg/kg) orally, one day before the operation. The surgical procedure was carried out under strictly aseptic conditions while the animals were anesthetized with isoflurane (induction 3.5% and maintenance 1.2% - 1.5%, at 0.8 L/min Oxygen). Throughout the surgical procedure, the animals received 5% dextrose in lactated Ringer's solution at a rate of 15 ml/kg/hr. Heart

rate, blood pressure and respiration were monitored constantly and recorded every 15 minutes. Body temperature was kept at 37.4° Celsius using a heating pad. Post-operatively, an opioid analgesic was administered (Buprenorphine hydrochloride 0.02 mg/kg, IM) every 6 hours for one day. Tylenol (10 mg/kg) and antibiotics (Tribrissen 30 mg/kg) were given to the animal for 3-5 days after the operation. At the end of the training period, another sterile surgery was performed to implant a *ball-and-socket* chamber for the electro-physiological recordings.

4.3.2 Visual Stimuli

Wire-like and spheroidal objects, similar to those used by Edelman and Bülthoff (1992) in human psychophysical experiments, were generated mathematically and presented on a color monitor (see insets in Figures 4-3, 4-8 and ??). All stimuli were presented on a monitor situated at a distance of 114cm from the animal.

The view generated by the selection of the appropriate parameters was arbitrarily named the *zero view* of the object. The viewpoint coordinates of the observer with respect to the object were defined as the longitude and the latitude of the eye on an imaginary sphere centered on the object. We used a right-handed coordinate system for the object transformations.

All objects were rendered using a visualization system (Application Visualization System, Stardent Computer Inc.) on a DEC5000 work station, and transferred to an IBM compatible AT486 computer(GATEWAY 486/33C). Display of the images was accomplished by means of a graphics card (Number Nine Computer, SGT board) of 640 x 480 resolution, at 60Hz refresh rate.

4.3.3 Animal Training

The monkeys were trained to recognize objects irrespective of position or orientation. They were first allowed to inspect an object, the *target*, presented from a given viewpoint, and subsequently were tested for recognizing views of the same object generated by $\pm 10°$ to $\pm 180°$ rotations around the vertical axis. The images were

presented sequentially, with target views dispersed among a large number of other objects, the *distractors* [2]. Two levers were attached to the front panel of the primate chair, and reinforcement was contingent upon pressing the right lever each time the target was presented. Pressing the left lever was required upon presentation of a distractor. Correct responses were rewarded with fruit-juice.

Initially, animals were trained to recognize the target's zero view among a large set of distractors, and subsequently they were trained to recognize additional target views resulting from progressively larger rotations around one axis. After a monkey learned to recognize a given object from any viewpoint in the range of $\pm 90°$, the procedure was repeated with a new object. Within an object class, the similarity of targets to distractors was gradually increased. A criterion of 95% correct for several objects was required in order to proceed with psychophysical data collection.

In the early stages of training the animals to perform the recognition task, a juice reward followed each correct response. As training progressed, the continuous reward paradigm was gradually replaced with a variable-ratio schedule. Finally, in the last stage of the training, the monkeys were rewarded only for ten consecutive correct responses. The end of the observation period was signaled by an increased juice reward and a green flash which filled the screen.

During the training, irrespective of reinforcement schedule, the monkeys always received feedback as to the correctness of their response since one incorrect report aborted the entire observation period. In contrast, no feedback was given to the monkeys during the psychophysical data collection. To discourage arbitrary performance, or the development of hand-preferences, *e.g.* giving only right hand responses, sessions of data collection were randomly interleaved with sessions with novel objects, in which incorrect responses aborted the trial.

[2]See Section 2.2

4.3.4 Task Description and Data Collection

The animals are trained to be able to perform two tasks: an active recognition task and a passive fixation one. The active recognition task has been used in the view-dependent experiment. The active part of the task is important so both behavioral data and cells activity could be collected at the same time. To recognize objects in the periphery is a difficult task for the monkeys. They tend to try to fc veate.

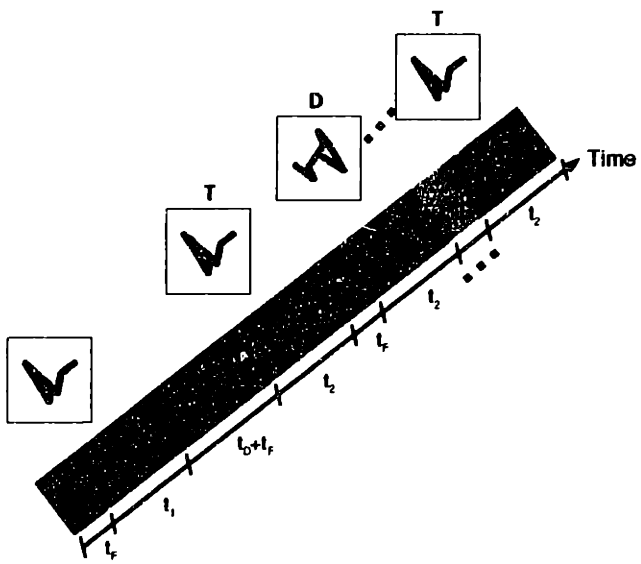

Figure 4-3: The recognition task. After a study phase in which the target object is presented as a motion sequence, a test phase followed where targets (T) and distractors (D) were presented in sequence. Each presentation required a response from the monkeys.

The sequence of events in a single observation period is described in Figure 4-3. Successful fixation is followed by the *training* or *learning phase.* In this phase, the target is inspected for 2 to 4 seconds from one or two viewpoints, called the *training views.* To provide the subject with $3D$ structure information, the target is presented as a motion sequence of 10 adjacent, shaded views, 2° apart, centered around the training view or views. The animation is accomplished at a two frames-per-view temporal rate, *i.e.* each view lasted 33.3 ms, yielding the impression of an object oscillating slowly $\pm 10°$ around a fixed axis.

The learning phase is followed by a short fixation period after which the *testing phase* begins. Each testing phase consists of up to 10 trials. The beginning of a

trial is indicated by a low-pitched tone, immediately followed by the presentation of the test stimulus, a shaded, static view of either the *target* or a *distractor*. Target views are generated by rotating the object around one of four axes, the vertical, the horizontal, the right oblique, or the left oblique. Distractors are other objects of a same or different class. The duration of stimulus presentation is 500-800 ms. The monkeys are given up to 1500 ms to respond by pressing one of the two levers attached in front of the primate chair. Typical reaction times are below 1000 ms for all the animals. An experimental session consists of a sequence of 60 observation periods, each of which lasts about 25 seconds.

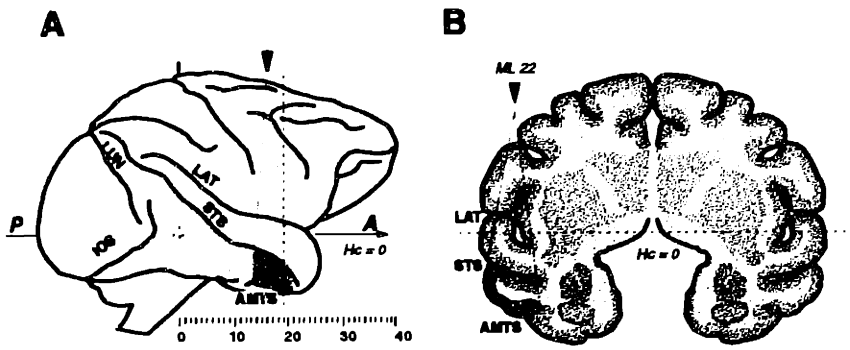

Figure 4-4: Anatomical location of the recording sites estimated from the stereotaxic coordinates. A. Lateral view of a macaque brain. The inferior occipital sulcus (IOS), the superior temporal sulcus (STS), and the anterior medial temporal sulcus (AMTS) are labeled. The dashed, vertical line marks 19mm anterior which is the location of the frontal section presented in B. The thin, gray lines represent the roughly conical volume accessible for recording using a ball-and-socket electrode drive. The anterior-posterior and medial-lateral extents of the primary recording site (dark gray and black) were from 14mm to 21mm anterior and 16mm to 24mm lateral. The black area denotes the estimated location in which the data presented in this paper were collected.

During each observation period, the animals' eye movements are measured using the scleral search coil technique (Robinson, 1963), and digitized at 200Hz rate. Manual responses are recorded at 200Hz through a digital I/O interface. Recording of single unit activity is done using Platinum-Iridium electrodes of 2-3 MΩ impedance. The electrodes are advanced into the brain through a guide tube mounted into a ball-and-socket positioner. By swiveling the guide tube, different sites can be accessed within an approximately 10x10mm cortical region (see Figure 4-4). Action potentials

are amplified (Bak Electronics, Model 1A-B), and routed to an audio-monitor (Grass AM-8), and also to a time-amplitude window discriminator (Bak Model DIS-1). The output of the window discriminator is used to trigger the real-time clock interface (KWV11) of the computer (PDP11/83).

4.4 Viewpoint-Dependent Recognition Performance

Monkeys were trained to recognize any given object viewed on one occasion in one orientation, when presented on a second occasion in a different orientation. Technically, this is a typical "old-new" recognition memory task during which the subject is required to state for each stimulus whether it is familiar (old, target), or unfamiliar (new, distractor). The probability of the subject reporting *familiar* when presented a target determines the *hit rate*, while the probability of reporting *familiar* when presented a distractor determines the *false alarm rate.*

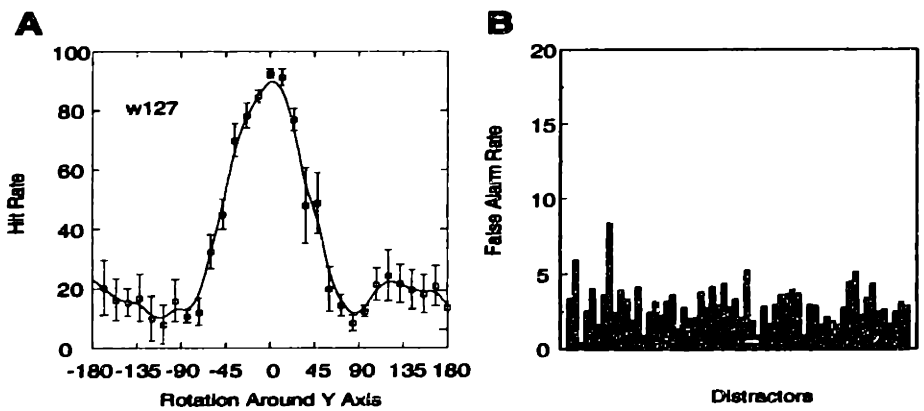

Figure 4-5: Recognition performance as a function of rotation in depth for a wire-like object.

Figure 4-5 shows the average performance of a monkey tested for the recognition of a wire-like object. Both the target and the distractors were generated using the same constraints, i.e. they had a similar moment of inertia, similar variability in segment orientation, and identical segment length and thickness. Thirty target views generated from rotations around the vertical axis, and 60 distractor objects were

used during testing. The small rectangles show mean performance for each tested view over 4 sessions of 40 presentations each. Two monkeys were tested, both for 4 sessions. The solid line was obtained by a distance weighted least squares smoothing of the data using the McLain algorithm (McLain, 1974). The two monkeys could correctly identify the views of the target around the trained, zero view, while their performance dropped below chance levels for rotations larger than $\pm 45°$. Performance below chance level is probably the result of the large number of distractors used within a session, which limited opportunity for learning of distractors *per se*. Therefore an object that was not perceived as a target view, was readily classified as a distractor. (Figure 4-5B shows the false alarm rate).

To exclude the possibility that the observed view dependency was specific to non-opaque structures lacking extended surface, we also tested recognition performance using spheroidal, amoeba-like objects with characteristic protrusions and concavities. As with the wire-like objects, the monkeys were able to generalize only for a limited number of novel views clustered around the view or views presented in the training phase (see Logothetis and Pauls, (1994)).

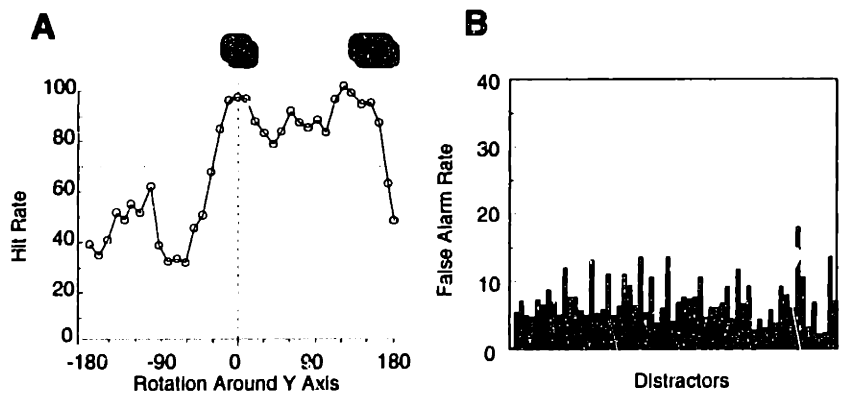

Figure 4-6: Interpolation between two familiar views

The ability to generalize recognition to novel views was also examined, after training the animals with two successively presented views of the target 75°, 120°, and 160° apart. Interpolation was found to be complete (above 95% performance) for training views 75° apart. Error rate increased for views 120° apart, and for a view disparity of

160° the monkeys were unable to interpolate recognition. Figure 4-6 shows the results of the experiment in which training was done with 2 views of a wire-like object, 120° apart.

During testing, the monkeys were first shown briefly the two familiar views of the object, and then ten stimuli in succession that could be either target or distractor views. The stimuli were pseudo-randomly selected from a set of 36 targets and 60 distractors. Within one experimental session, each of the 36 tested target views was presented 30 times. As can be seen in Figure 4-6, the performance of the animal remains above criterion (75%) for all views between and around the trained views. Figure 4-6B shows the false alarm rate for the distractors tested. Training with 3 to 5 views proved enough for generalizing around one view circle.

Performance was found to be viewpoint-invariant when animals were tested for basic level classifications, or when they were trained with multiple views of wire-like or amoeba-like objects. It should be noted that the term "basic-level" is used here to simply denote that the target objects were largely different in shape from the distractors. Distractors, in this case, were selected from a set of 120 objects, including geometrical constructs, altered wires or spheroidals, plane-models, or fractal objects. Since all animals were already trained to perform the task, independent of the object type used as a target, no familiarization with the object's zero-view preceded data collection in those experiments. Nonetheless, the animals generalized recognition for all tested novel views

4.4.1 Is viewpoint-dependent performance due to a failure to understand the task?

We have provided evidence for a similar view-dependency of recognition in the non-human primate. Monkeys were indeed unable to recognize objects rotated more than approximately 40° of visual angle from a familiar view. These results are hard to reconcile with theories postulating object-centered representations. Such theories predict uniform performance across different object views, provided $3D$ information

is available to the subject at the time of the first encounter. Therefore, one question calling for discussion is whether or not information about the object's structure was available to the monkeys during the learning phase of these experiments.

Experiments on monkeys have shown that non-human primates, too, possess the ability to see structure from motion (Siegel and Andersen, 1988) in random-dot kinematograms. Thus, during the learning phase of each observation period, information about the 3D structure of the target was available to the monkeys by virtue of shading, the kinetic depth effect, and minimal self-occlusion.

Could the view-dependent behavior of the animals be a result of the monkeys' failing to understand the task? They could indeed recognize a two-dimensional pattern as such, without necessarily perceiving it as a view of an object. Correct performance around the familiar view could then be e plained as the inability of the animal to discriminate adjacent views. Several lines of arguments refute such an interpretation of the obtained results. First, the animals easily generalized recognition to all novel views of common objects. Moreover, when the wire-like objects had prominent characteristics, such as one or more sharp angles, or a closure, the monkeys were able to perform in a view-invariant fashion. Second, when two views of the target were presented in the training phase the animals interpolated, often with 100% performance, for any view between the two trained views. Finally, human subjects that were tested for comparison using the same apparatus, exhibited recognition performance very similar to that of the tested monkeys.

Thus, it appears that monkeys, just like human subjects, show rotational invariance for familiar, basic-level objects, but they fail to generalize recognition at the subordinate level, when fine, shape-based discriminations are required to recognize an object. Interestingly, training with a limited number of views (about 10 views for the entire viewing sphere) was sufficient for all the monkeys tested to achieve view-independent performance.

Recognition based entirely on fine shape discriminations is not uncommon in daily life. We are certainly able to recognize modern sculptures, mountains, or cloud formations. The largely view-independent, basic-level recognition exhibited by adults may

be the result of learning certain irreducible shapes early in life. Even those theories suggesting that recognition involves the indexing of a limited number of volumetric components (Biederman, 1987) and the detection of their relationships have to face the problem of learning components that cannot be further decomposed. In other words, we still have to achieve representations of some elementary object forms that transcend the special viewpoint of the observer. Such representations usually rely on shape coding that is very similar to what is required for the subordinate level of recognition.

4.5 View Selectivity in Inferior Temporal Cortex

Cells selective for specific patterns or object views are not rare throughout the IT cortex. View selectivity has been reported previously for face-selective IT neurons. Desimone et al (1984) reported cells that were sensitive to the orientation of the head in depth. In their example, the cell's activity fell to half of its maximum when the face was rotated about 30° to 40°, which is in close agreement with the data we presented using the wire-like or amoeba objects.

In the present study, we found neurons that responded selectively to novel visual objects that the monkeys learned to recognize during the experiments. None of these objects had any prior meaning to the animal, and none of them resembled anything familiar in the monkeys' environment. Thus it appears that neurons in this area can develop a complex receptive field organization as a result of extensive training in the discrimination and recognition of objects.

The monkeys were trained with different types of objects, such as the wire-like, the spheroidal, and basic types of objects. Interestingly, the frequency of encountering neurons selective to a particular object-type seemed to be related to the animal's familiarity with the object class. In one of the monkeys, the wire-like objects, extensively used during the psychophysical experiments, were much more likely to elicit cell responses (71 selective cells) than, for example, spheroidal objects (10 selective cells), which were used to a much lesser extent. The converse was observed in another

animal that was extensively trained with the amoeba-like objects. In the case of both object types, most selective neurons responded best to one view of the object, while their response decreased as the object was rotated away from the preferred view. Plotting the cell responses as a function of rotation angle revealed systematic view tuning curves similar to those obtained from striate neurons tested with lines rotated in the frontal plane.

The activity of neurons in the IT cortex was examined in a simple fixation task, and in the experimental paradigm described earlier. We collected data from over 970 isolated IT neurons in two macaque monkeys. Figure 4-4 shows the recording sites as estimated from the stereotaxic coordinates. Since the animals are being used in further experiments on object recognition, no histological reconstructions are currently available.

Isolated units were tested with a variety of simple or complex patterns while the animal was involved in a fixation task. The animal was trained to maintain fixation within a $1°x1°$ window. The activity of cells that responded either to the wire or the amoeboid objects was further examined while the animal performed the recognition task.

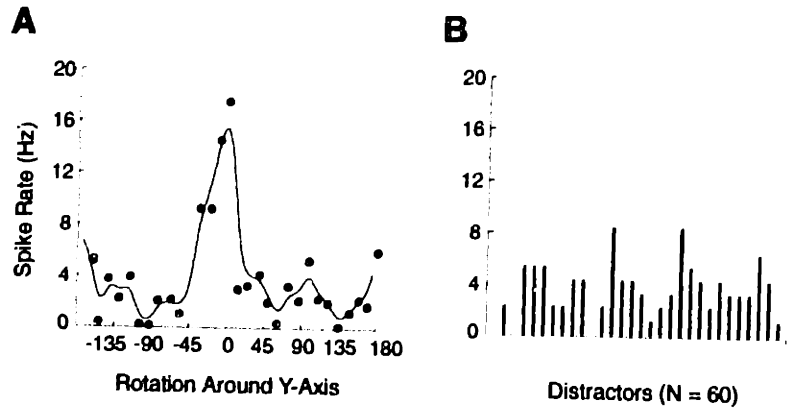

Figure 4-7: Cell response to targets and distractors

We found a number of units showing a remarkable selectivity for views of individual wire objects that the monkeys had learned to recognize. An example of a view-selective neuron is shown in Figure 4-7A. The responses of this particular cell

were studied for 30 different views of a wire-like object that the monkeys were trained to recognize. The monkeys performance was above 95% for all the views of this object. The cell was highly selective for views located around 0°. Its activity decreased considerably with even a 12° deviation from the preferred view. Individual wire segments with the same orientation as the segments of the preferred view did not elicit any response when presented in the same location of the receptive field. Figure 4-7B shows the responses of the same cell for distractor objects.

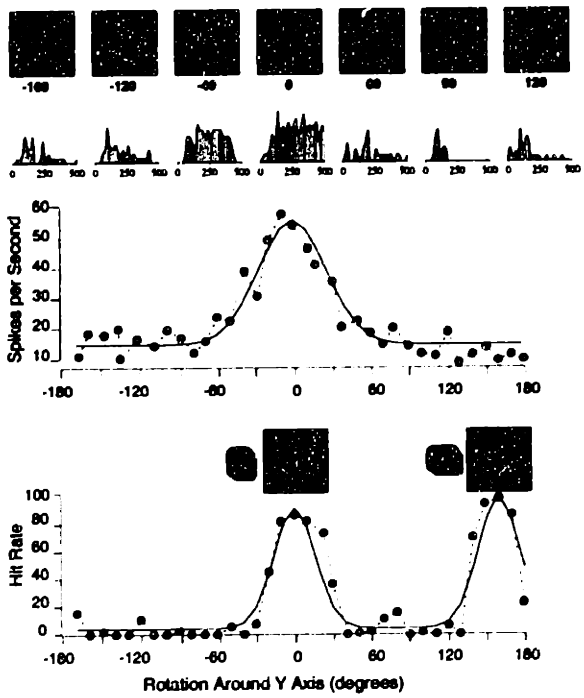

Figure 4-8: View-selective cell response and behavioral performance

Figure 4-8 shows a second example of a view-tuned neuron. The monkey was trained to recognize the views labeled 0° and 160° of the wire object, and subsequently was tested for 36 views around the same axis with no feedback as to the correctness of responses. As mentioned above, when the training views were this far apart, no interpolation was observed during the testing period (see lower plot in Figure 4-8). The monkey appears to learn the two views in a non-associative manner, just like learning two different wire-objects. The cell responded for views around the zero view of the object, but it did not discharge for the other training view.

Many other cells showed a tuning that was likewise independent of the animal's

behavioral performance. These observations provide strong evidence that the activity of these neurons is not simply the manifestation of arousal, attention, or a sensation of familiarity, but rather it relates to the object's characteristic features or views. To date, 71 out of the 970 (6%) analyzed cells showed view selective responses similar to those illustrated in Figure 4-8. In their majority, the rest of the neurons were visually active when plotted with other simple or complex stimuli, including faces. A small percentage of neurons, although firing frequently with a rate of 5 to 20Hz, could not be driven by any of the stimuli used in these experiments.

A small percentage of cells (8 out of 970) responded to wire-like objects presented from any viewpoint, thereby showing view-invariant response characteristics. No selective responses were ever encountered for views that the animal systematically failed to recognize. Finally, for three objects, more than one neuron was found to be tuned to different views of the same object.

4.5.1 Invariance for Reflections

Some of the neurons presented in teh previuos section showed similar response magnitudes for views of the wire-like objects that were 180° apart. As mentioned earlier, these views tend to look like mirror-symmetrical images of two-dimensional patterns, thus implying some sort of "reflection-invariance" in the response of IT neurons.

Interestingly, some face cells have been found to respond to views of a face 180° apart, especially the left and right profiles. These cells are presumably similar to those reported here as responding to the "pseudo-mirror-symmetrical" views. Reflections invariance has also been shown in the responses of some face selective IT cells in infant monkeys, a finding that suggests that such invariance may be generated automatically for every learned object, and may be present already early in an individual's life.

In fact, such an invariance may be the cause of the inability of children to distinguish between mirror-symmetrical letters like 'd' and 'b'. This type of letter confusion was studied intensively in children by Orton (1928). He observed a delay in the learning of mirror-symmetrical letters and words as well as several other characteristics involving the establishment of "handedness" in language handicapped children. This

eventually lead to the description by Orton of a disorder known as *strephosymbolia.* This confusion, observed in normal children as well, appears to be the rule during development and not the exception. In fact, Gross and Bronstein suggest that the confusion of mirror images may be an adaptive mode of processing visual information rather than a real "confusion" (?). These authors note that, in the natural world, there are never any mirror images that would be useful for an animal to distinguish. For example, in the case of bilateral symmetry observed in most animals, the two mirror-symmetrical sides are aspects of the same thing, and it would be more adaptive to treat them the same.

4.6 Responses of IT neurons to scaling and translation of novel 3D objects

As described in the literature, and confirmed for our stimuli by preliminary experiments, cells in the anterior part of IT are considered to be translation invariant across few degrees. If we carefully look at the data reported in the literature, we see that IT cells have large receptive fields that almost always include the fovea, and the highest response is in the fovea region. In the case of shape selective cells, selectivity is maintained over changes in stimulus position and size (Gross and Mishkin, 1977). Schwartz et al. reported that absolute firing rate, but relative preference, for a particular stimulus is maintained (Schwartz et al., 1983).

View-selective neurons were also tested with respect to changes in position or size of the preferred view. The task used to test these transformations was identical to that described above for all sizes and for small positional displacements. The stimulus sizes used subtended from 1.9 to 5.6 degrees of visual angle. Four eccentric positions were tested all at a distance of 2.25 degrees (the result of a translation of the object's center 2.25 degrees, roughly 1 object radius, from the center of gaze). For larger translations the animal was only required to fixate without performing the recognition task. An example of a view-selective neuron responding invariantly to changes in size but not position is shown in Figure 4-9. This particular cell was

selective for a limited region of the object around $-108°$ (Figure 4-9a), and responded more that 10 times more for the preferred target view than for the best distractor (Figure 4-9b).

Figure 4-9c shows the ratio of the target response to the mean response for the ten best distractors for the sizes tested. Note that all of the distractors were of the default size and were presented foveally. The equality in the height of the bars indicates that the cell maintained both its selecitivity and its mean firing rate for the preferred stimulus over changes in size. The responses of the same cell to translation are plotted in Figure 4-9d. This particular neuron showed some variance in its response depending on stimulus position, however, in all cases its response was for an eccentrically presented target was still at least 2.5 times that for foveally presented distractors. We found seventy-five percent of the view-selective cells tested to be invariant to changes in stimulus size. Responses, however, varied according to the position of the stimulus in the receptive field, with stronger responses usually elicited in the foveal region. Only about thirty-five percent of the cells tested were invariant to changes in stimulus position for the limited range of translation tested. In our hands, the responses of IT neurons lie on a continuum composed of cells invariant to either scale, position or both to varying degrees.

4.7 Discussion and Conclusion

Our results provide evidence supporting a viewer-centered representation of objects in the primate, at least for subordinate level classifications. The main findings of this study can be summarized as follows:

1. Even when complete information about the structure of an object is available to the subject, recognition at the subordinate level depends on the object's attitude, both for monkeys and for human subjects.

2. A memory-based, viewer-centered recognition system is not an implausible mechanism for object-constancy. Both theoretical work, and the results pre-

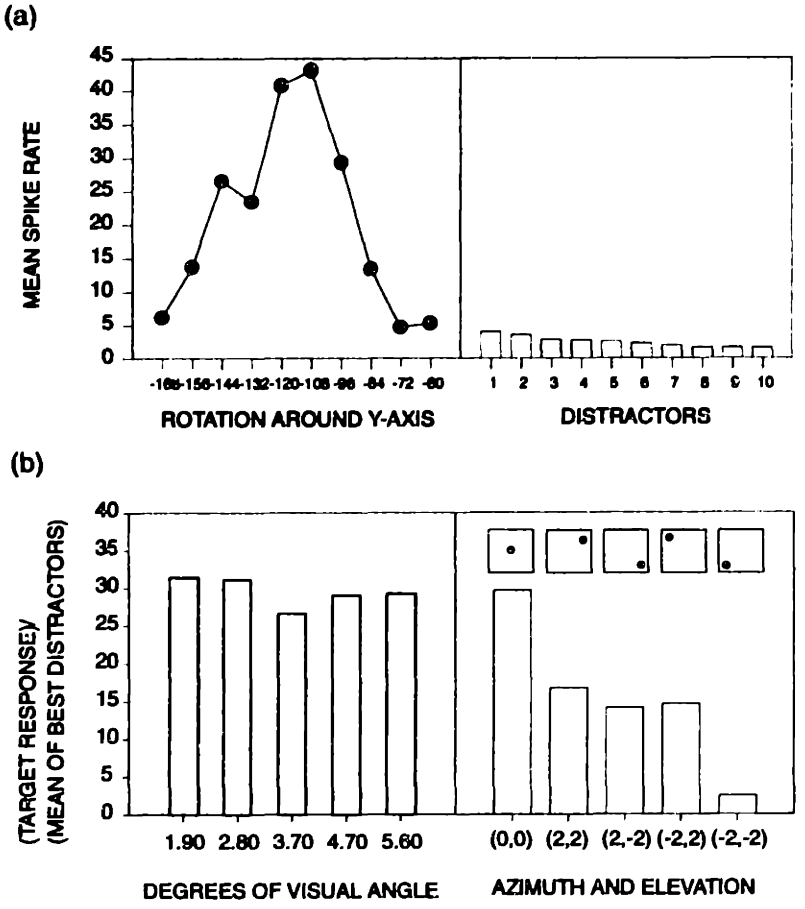

Figure 4-9: Responses of a view-selective IT neuron to scaling and translation of the preferred view.

sented here, suggest that only a small number of object views need to be stored in order to achieve perceptual invariance. The ability of both humans and monkeys to interpolate between two familiar views depends on their distance, a finding difficult to reconcile with a recognition system based on linear interpolation (Ullman and Basri, 1991), but directly predicted by a system relying on nonlinear approximation (Poggio and Girosi, 1990b).

3. A small population of IT neurons has been found to respond selectively to individual members of the object-classes tested in this study. The response of some neurons is a function of the object's view. The discharge rate of many IT neurons is found to be a bell-shaped function of orientation, centered on a preferred view.

4. For all objects used in the combined psychophysical-electrophysiological experiments, view-tuning is observed only for those views that the monkeys can recognize. We find several neurons that respond to the sight of unfamiliar or distractor objects. Such cells, however, give non-specific responses to a variety of other patterns presented while the monkeys performed a simple fixation task.

5. A subset of the view-selective units were also tested for position and size changes of the preferred view. In our experiments, the responses of IT neurons lie on a continuum from cells invariant to either scale, position, or both, to cells with varying degrees of sensitivity to these parameters.

Thus, the psychophysical performance of both humans and non-human primates seems to be consistent with the idea that a network founded on view-based approximation modules synthesized during training, such as the RBF network model, may indeed be one of several algorithms the primate visual system uses for object recognition. Sets of neurons tuned broadly to individual object views may represent the neural substrate of such approximation modules. A very small number of neurons exhibit object-specific but view-invariant responses, which might be the result of the convergence of view-dependent units into neurons showing characteristics of object-centered descriptions. The input of each view-selective unit can be considered as the

conjunction of simpler features extracted at earlier stages in the visual system. Such a scheme is obviously oversimplified and lacks top-down mechanisms that strongly affect recognition performance. The processing of object information is undoubtedly far more complex, and representations might be local and explicit, or distributed and implicit, according to the recognition task or the stimulus context. Although the ultimate goal of a recognition system is to describe grouped object-features in a more abstract format that captures the invariant, 3D, geometric properties of an object, early representations may be strongly configurational in some cases. Moreover, for visually complex, non-decomposable objects, like many biologically-meaningful objects, only holistic representations may be the only ones possible. Neurons selective for object-views and tolerant of varying extents of image transformations may then be elements of one possible mechanism for such representation.

Chapter 5

Analysis of the view dependence of population codes in inferior temporal cortex

The discovery in inferior temporal (IT) cortex of cells showing responses selective for specific views of learned wire-like objects (Logothetis et al., 1995) supports a computational model of recognition. Such a model is built on two-dimensional view-based representations (Vetter et al., 1995). However, many more IT cells show significantly view-sensitive responses without strict view-tuning, which suggests that an analysis of population encodings may be necessary fully to understand the nature of object representation in IT. We analyze the dependence of similarity of IT population responses to views of objects obtained by rotating in depth, on the orientation difference between the views. This provides further evidence that neural representations of object shape depend on abstract two-dimensional views (possibly built from collections of image features). With such an analysis, we also demonstrate a clear correspondence between the view dependence of IT population responses, and the view dependence of human observers' similarity judgments.

The primate visual system can effortlessly recognize an innumerable variety of

[1]This work was done in collaboration with Josh Tenenbaum and appears in Tenenbaum and Bricolo, 1996

objects. Recognition performance is robust to common imaging transformations that preserve object identity: a particular chair can still be identified easily when viewed from different viewpoints, at different distances, in different positions and under different illumination conditions. Over the last two decades, lesion experiments and electrophysiological recordings have established the crucial role played by the IT cortex in object recognition (Gross, 1994). Specifically, the recent discovery by Logothetis and colleagues (Logothetis et al., 1995) that a significant number of IT cells show responses selective for specific views of wire-like objects learned while performing a recognition task (see Figure 5-1a) has provoked strong interest in computation and psychophysics circles (Bülthoff et al., 1995). As we have already noted, this finding offers some support for computational models of recognition, built on two-dimensional view-based representations (Vetter et al., 1995) over models based on three-dimensional structural descriptions (Biederman, 1987), and thus complements the substantial body of psychophysical evidence indicating that human object recognition is significantly view-dependent (Bülthoff et al., 1995).

However, two considerations should keep us from drawing any definitive conclusions about the nature of object representations in IT from the mere presence of view-selective cells. First, only 7.7% of the IT cells analyzed in Logothetis and Pauls (1995) responded significantly better around one or two views of a target object, relative to other views of the target, and also to the other tested distractor objects. Second, much more common than view-tuned cells (Figure 5-1a), are cells like those shown in Figure 5-1b-d. These latter cells, like the view-tuned cells, show significantly stronger responses to some views of the target object than to other views, but do not have one or two sharply preferred object orientations, and do not necessarily respond significantly better to the target object than to all distractors. Evidently, much of the IT cell population finds the variation in views of the target object produced by rotation in depth meaningful, but only the minority of these view-sensitive cells seem interpretable under the standard paradigm of view-based representations. In this section, we analyze the object representations encoded in IT population responses, including mainly cells that are more or less view-sensitive but not strictly view-tuned,

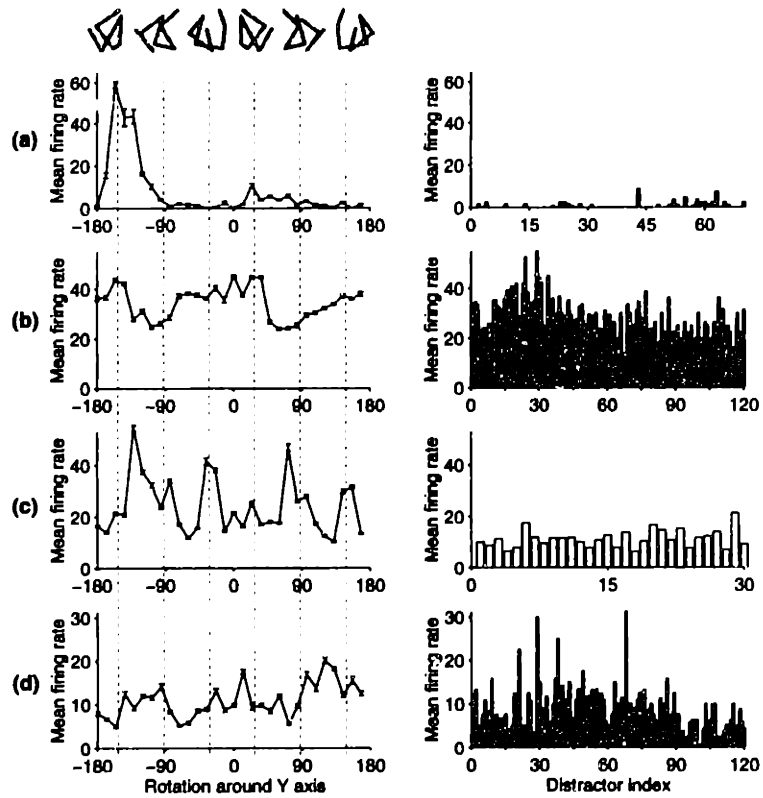

Figure 5-1: The response (mean firing rate) of four different IT cells to 30 views of a wire-like object rotated in depth (first column), and to various views of other distractor objects (second column).

using the recordings originally reported and analyzed at the single-cell level in (Logothetis et al., 1995). We show that population representations are most consistent with theories of object recognition based on two-dimensional abstract views, paralleling the findings of recent computational and psychophysical work (Vetter et al., 1995; Bülthoff et al., 1995). We also demonstrate a clear correspondence between the view dependence of IT population responses, and the view dependence of human observers' similarity judgments.

5.1 Analytical methods

The activity of IT cells was recorded while monkeys performed a visual discrimination task (see Chapter II and Logothetis and Pauls, 1995 for details). In the case of two of the wire objects used as target stimuli in this task (see Figure ??), more than 60 cells were recorded with significant responses. We chose to study these two objects, and the corresponding cell populations[2] ($N_1 = 63$ cells and $N_2 = 70$ cells), since they provide the closest approximations to the complete IT population code.

We have focused on analytic techniques for population codes, with the potential to discriminate directly between competing theories of object representation for recognition. In this section, we compare the similarity of the neural population responses to the 30 views of each object, which are generated as the object rotates in depth about a central vertical axis, with the predicted similarities of the representations of these 30 views, as offered by four candidate theories of object representation (Bülthoff et al., 1995; Pinker, 1984): (i) three-dimensional orientation-invariant (which ignores the object's 3-D pose); (ii) three-dimensional orientation-dependent (which explicitly encodes the object's 3-D pose); (iii) two-dimensional feature-based (composed of combinations of image features); (iv) two-dimensional template-based (composed of essentially image-like arrays).

In general terms, these four representational schemes are ordered by their degree of geometric abstraction, and consequent invariance under the nonlinear imaging transformation of rotation in depth. Thus, they yield different predictions as to how the similarity of two views may depend on the orientation difference between them. The 3-D orientation-invariant scheme predicts no effect of orientation difference on view similarity, because information about the object's 3-D pose has already been abstracted away from this representation. The 3-D orientation-dependent scheme predicts that view similarity is simply a monotonically-decreasing function of orientation difference. Both 2-D schemes also predict that view similarity generally decreases with orientation difference. However, the 2-D template-based scheme pre-

[2]Both populations were collected from the same monkey

dicts that orientation difference alone can account for only a small percentage of the variance in view similarity, because rotation in depth produces very nonlinear image deformations. The 2-D feature-based scheme predicts that orientation difference can account for much of the variance in view similarity, because key image features, such as vertex angle or segment length in the case of wire-like objects, change smoothly as the object rotates in depth, but under this scheme, view similarity may not always decrease monotonically with increasing orientation difference if the key features are visibly indistinguishable from more than one viewpoint (e.g. left and right, front and back).

Following previous similarity analyses of neural population codes (Young and Yamane, 1992b; Weiss and Edleman, 1995), we treat the population response to each view as a vector in a high dimensional "cell space", with each component dimension corresponding to the mean firing rate of one cell in the population recorded while the monkey is presented with that view. Firing rates are normalized for each cell by the overall mean firing rate of that cell to all 30 views of the same object. We then measure the population similarity of two views by taking the negative Euclidean distance in cell space, with the contribution of each cell's response weighted by its standard error, and scale these values uniformly to fill the interval $[0, 1]$. Here we present the results of two analytical procedures applied to the computed view similarities. First, in order to assess the level of representational abstraction in the population code for each object, we use multidimensional scaling (MDS) to represent the view similarities as distances in a Euclidean space of two dimensions (the minimum dimensionality needed to accommodate the "circle" of views produced by rotation in depth). Then we calculate the percentage of variance in the complete view similarities accounted for by these abstract, reduced-dimensionality view spaces (Shepard and Farrell, 1985; Young and Yamane, 1992b; Weiss and Edleman, 1995). Now, in order to uncover the specific relation between view similarity and the orientation difference between views, we average the similarity of all pairs of views, separated by each orientation difference, in order to obtain a single similarity gradient for each object (Figure 5-2).

5.2 Results and Discussion

For the two objects, two-dimensional configurations of points produced by MDS could account for (respectively) 72.7% and 68.4% of the variance in the view similarities originally computed from the IT populations in (respectively) 63-dimensional and 70-dimensional cell spaces. Thus, the variation in population responses due to rotation in depth is not strictly reducible to two Euclidean dimensions, as would be expected with a 3-D orientation-dependent representational scheme. On the other hand, the variances accounted for by these two-dimensional configurations are much greater than the values of 39.7% and 32.8% predicted by a 2-D template-based representational scheme, in which view similarity is computed as normalized image pixel correlation. These MDS results thus suggest that IT populations may encode 2-D feature-based representations of objects, at an intermediate level of abstraction between 2-D image template-based representations and a 3-D orientation-dependent representations. As shown below, a more detailed analysis of the relation between view similarity, and the orientation difference between views, confirms this hypothesis.

The left-hand panels of Figure 5-2a-b show the empirically obtained relation between view similarity and orientation difference, as expressed by IT population codes for the two target objects. Both curves show significant violations of monotonicity: view similarity decreases with increasing orientation difference, but only for differences up to 90°; subsequently, similarity increases again as orientation difference increases to a second peak at 180°. Note that the shape of these curves is not significantly altered by excluding from the analysis those few cells that are distinctly tuned for two views 180° apart.

For the sake of comparison, the central panels of Figure 5-2 show the relation between view similarity and orientation difference as predicted by a 2-D template-based representational scheme. Each point on these curves represents the normalized image pixel correlation of all pairs of views separated by the indicated orientation difference. As expected, there is no hint of a significant departure from monotonicity. Clearly, a 3-D orientation-dependent representation would give the same monotonically decreas-

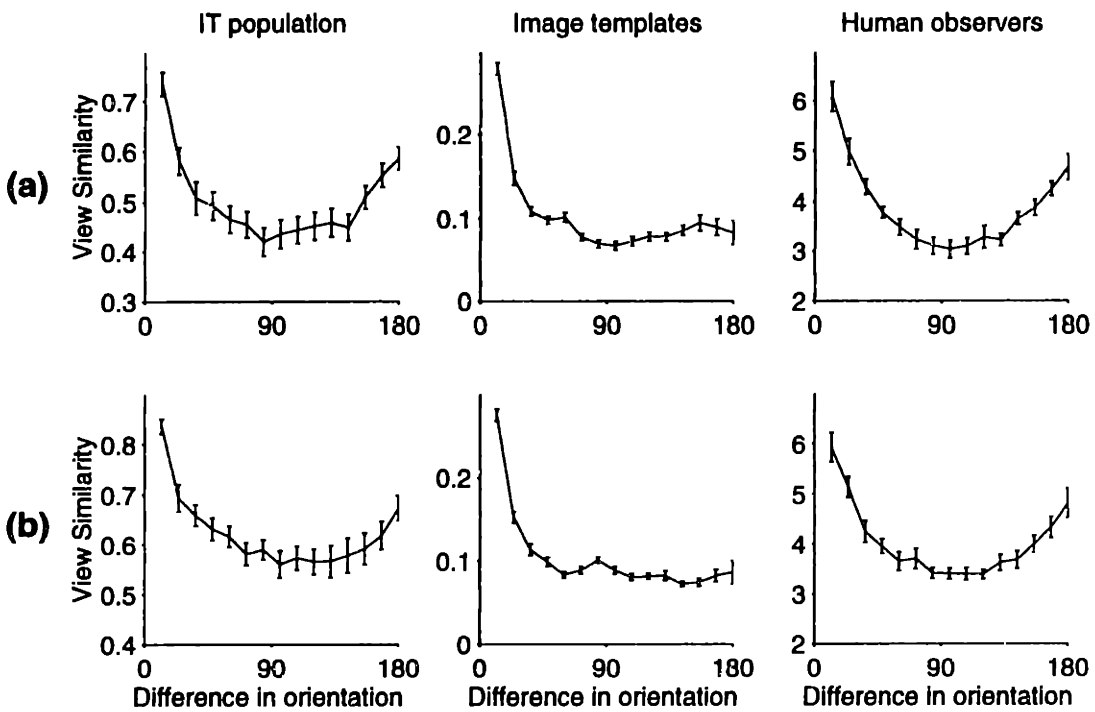

Figure 5-2: The relation between view similarity and orientation difference for two objects (shown in (a) and (b) respectively), as calculated from IT population codes, an image template correlation model, and the judgments of human observers.

ing relation, and a 3-D orientation-invariant representation would generate completely flat curves. Thus, of the four representational schemes discussed in the previous section, the empirical U-shaped curves are most consistent with abstract 2-D view-based representations built from collections of image features, such as vertex angles or segment length, that appear indistinguishable in two views separated by 180°. Single view-tuned cells may be selective for combinations of these features, as well as for other image features, such as segment direction or ordinal segment occlusion relations, that are not invariant under rotation by 180° and hence may account for why the peak at 180° is still much lower than the maximum.

Of course we don't always expect to obtain such dramatic U-shaped curves for natural objects, with features that are often occluded by large rotations in depth. However, natural objects are often bilaterally symmetric, and both computational

considerations (Vetter et al., 1995) and psychophysical findings (Vetter et al., 1994) lead us to expect an approximately U-shaped relation between view similarity and orientation difference to hold for these objects as well.

The rightmost panels of Figure 5-2 show the relation between view similarity and orientation difference, according to the judgments of human observers. We asked 10 subjects to rate the similarity of each pair of views of the same two wire objects on a 1-7 scale, and averaged their judgments so as to obtain these curves. The curves clearly exhibit the same U-shape which we obtained from the IT population responses. Given that object recognition depends so heavily on shape similarity, this provides further support for the hypothesis that view-dependent recognition performance in psychophysical tasks results from the particular view-based representations of objects encoded in IT cortex.

Part III

Computational models

Chapter 6

A model of view-tuned units

6.1 Introduction

Among the many abilities of biological systems, recognition of visually-presented objects is one of the most remarkable. Familiar objects can be readily recognized from their shape, color, or texture. Even when partially occluded behind another surface, they can be "guessed", based on some sort of reasoning processes driven by contextual information. Moreover, an object presented briefly on a single occasion can, as a rule, be perceived and recognized on a second occasion, even when seen from a different viewing distance, and rotated in the frontal plane. Thus, it appears that the visual system can not only form new internal representations of objects very quickly, but it can also access these representations through entirely different input configurations.

In striking contrast, object recognition has proved to be very difficult to achieve in artificial systems. This is partly because we know very little of what constitutes an object. There is nothing special about objects, at least not in the way they are represented in the input of the visual system. The shape of an object, or its "characteristic regions" or features, are almost never visually primitive constructions, determined by a predictable combination of primary cues. Any given $2D$ image can be parsed into an arbitrary set of objects, each of which can be recursively decomposed into smaller objects. What we consider to be an object depends on the visual input, yet it is also determined by the task at hand. The neural code of objects

is a mystery even when considering simple geometrical forms, such as a cube, a cone, or a cylinder, seen in isolation. Theories of object recognition posit that the visual system stores a representation of an object, and that recognition occurs when the stored representation is matched to its corresponding sensory representation, when generated from the viewed object. This assumption, however, raises two obvious questions: what are these representations, and how is matching achieved? In the following chapter we will address the question of representation, and suggest a model that tries to capture general properties of neural object representation.

6.1.1 Object Representation Systems

With respect to representation, a major distinction is usually made between *object-centered* and *viewer-centered* spatial reference frames, within which the object features can be defined. The former is assumed to be intrinsic to the object, and thus implies a viewpoint-independent or invariant object representation, while the latter is defined with respect to the environment or the observer, and therefore predicts viewpoint-dependent recognition performance.

Human behavior is consistent with object-centered representations for the recognition of familiar objects. However, a serious problem arises from the fact that familiarity of the object implies that a recognition system based on $3D$ descriptions cannot easily be discerned from a viewer-centered system exposed to a sufficient number of object views. Moreover, object-centered representations fail to account for human performance in studies of subordinate classifications with various kinds of novel objects. Indeed, different investigators (Rock and DiVita, 1987; Tarr and Pinker, 1989; Edelman and Bülthoff, 1992) have shown that human performance can be abysmal when subjects are asked to recognize novel views of some nonsensical and unfamiliar objects like the "cube", "wire", or "amoeboid" objects used by Tarr, and Edelman and Bülthoff respectively[1].

Viewer-centered representations, on the other hand, can account for recognition

[1] For an overview see Section 1.1

performance at any taxonomic level, but they are often considered implausible due to the vast amount of memory required to store all discriminable object views needed to achieve viewpoint invariance. However, a mathematical theory developed by Poggio and Girosi (Poggio and Girosi, 1990b), in combination with modeling work by Poggio and Edelman (Poggio and Edelman, 1990), demonstrates that viewpoint invariance may be achieved by a recognition system that can generalize from only a small number of familiar views.

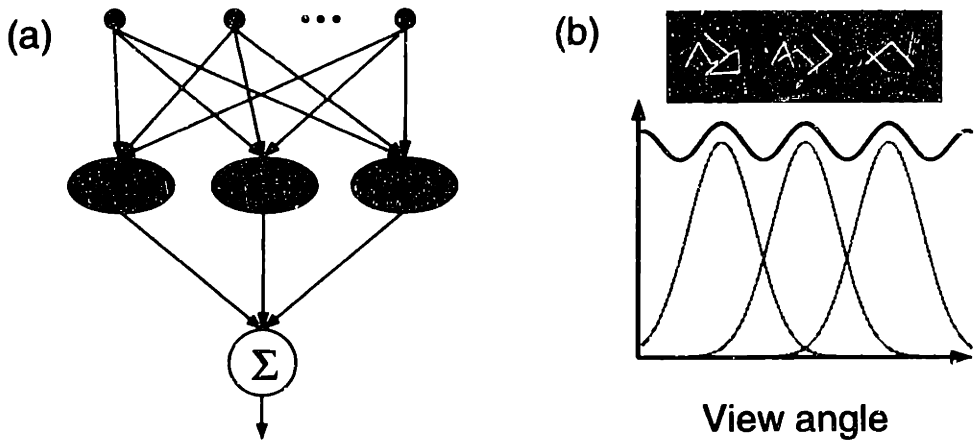

Figure 6-1: (a) Schematic representation of the architecture of the Poggio-Edelman model. The shaded circles correspond to the view-tuned units, each tuned to a view of the object, while the open circle correspond to the view-invariant, object specific output unit. (b) Tuning curves of the view-tuned (gray) and view-invariant (black) units.

6.1.2 Radial Basis Functions Networks

Poggio and Edelman (1990) have shown that a simple network can achieve viewpoint invariance by interpolating between a small number of stored views (Poggio and Edelman, 1990). Computationally, this network uses a small set of sparse data corresponding to an object's training views, in order to synthesize an approximation to a multivariate function representing the object (see Figure 6-1). In this sense, generalization of recognition is equivalent to the problem of reconstructing a surface from a sparse set of data points (Poggio and Girosi, 1990b; Poggio and Girosi, 1990a), and

can be approached in the framework of regularization theory. The problem of surface reconstruction is stated as follows: *suppose that the set* $g = \{(\mathbf{x}_i, y_i) \in R^d \times R\}_{i=1}^N$ *of data has been obtained by random sampling a function* f*, belonging to some space of functions defined on* R^d*, in the presence of noise: recover the function* f *given the set* g. In our case, the data points $\mathbf{x}_i$ are feature vectors representing the views of a specific object, and the corresponding output values y_i are labels (for example 1 and 0) that tells us whether the view $\mathbf{x}_i$ corresponds to that object or not. Therefore the function f is the characteristic function of a specific object, which has value 1 when the input is a view of that object and value 0 otherwise.

The problem of surface reconstruction is clearly ill-posed, since it has an infinite number of solutions. In order to choose one particular solution, we need to have some *a priori* knowledge of the function which must be reconstructed. The most common form of *a priori* knowledge consists in assuming that the function is *smooth*, in the sense that two similar inputs correspond to two similar outputs. The main idea underlying regularization theory is that the solution of an ill-posed problem can be obtained from a variational principle, which contains both the data, and the prior smoothness information. Smoothness is taken into account by defining a *smoothness functional* $\phi[f]$ in such a way that lower values of the functional correspond to smoother functions. Since we look for a function that is simultaneously close to the data and also smooth, regularization theory chooses as a solution of the approximation problem the function that minimizes the following functional:

$$H[f] = \sum_{i=1}^{N} (f(\mathbf{x}_i) - y_i)^2 + \lambda \phi[f] \tag{6.1}$$

where λ is a positive number that is usually called the *regularization parameter*. The first term is enforcing closeness to the data, and the second smoothness, while the regularization parameter controls the tradeoff between these two terms, and can be chosen, for example, according to cross-validation techniques (Craven and Wahba, 1979; Wahba, 1990).

It can be shown that, for a wide class of functionals ϕ, the solutions of the mini-

mization of the functional (6.1) all have the same form. In this section, we only report the main result, while a complete derivation can be found in Girosi et al. (1995) .

In order to understand the meaning of their result, we need first to give a more precise definition of what is meant by smoothness. We refer to smoothness as a measure of the "oscillatory" behavior of a function. Therefore, within a class of differentiable functions, one function will be said to be smoother than another one if it oscillates less. If we look at the functions in the frequency domain, we may say that a function is smoother than another one if it has less energy at high frequency (smaller bandwidth). The high frequency content of a function can be measured by first high-pass filtering the function, and then measuring the power, that is the L_2 norm, of the result. In formulas, this suggests defining smoothness functionals of the form:

$$\phi[f] = \int_{R^d} ds \, \frac{|\tilde{f}(\mathbf{s})|^2}{\tilde{G}(\mathbf{s})} \tag{6.2}$$

where $\tilde{}$ indicates the Fourier transform, $\tilde{G}$ is some symmetric positive function that tends to zero as $\|\mathbf{s}\| \to \infty$ (so that $\frac{1}{\tilde{G}}$ is an high-pass filter) and for which the class of functions such that this expression is well-defined is not empty. Under certain conditions on G, Girosi et al. showed that the function that minimizes the functional (6.1) has the form:

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i G(\mathbf{x} - \mathbf{x}_i) + \sum_{\alpha=1}^{k} d_\alpha \psi_\alpha(\mathbf{x}) \tag{6.3}$$

where $\{\psi_\alpha\}_{\alpha=1}^{k}$ is a basis in the k dimensional null space of the functional ϕ, that in most cases is a set of polynomials, and therefore will be referred to as the "polynomial term" in equation (6.3) and will be always indicate as $p(\mathbf{x})$. The coefficients d_α and c_i depend on the data, and satisfy the following linear system:

$$(G + \lambda I)\mathbf{c} + \Psi^T \mathbf{d} = \mathbf{y} \tag{6.4}$$

$$\Psi \mathbf{c} = 0 \tag{6.5}$$

where I is the identity matrix, and we have defined

$$(\mathbf{y})_i = y_i \ , \quad (\mathbf{c})_i = c_i \ , \quad (\mathbf{d})_i = d_i \ ,$$

$$(G)_{ij} = G(\mathbf{x}_i - \mathbf{x}_j) \ , \quad (\Psi)_{\alpha i} = \psi_\alpha(\mathbf{x}_i)$$

The approximation scheme of equation (6.3) has a simple interpretation in terms of a network with one layer of hidden units, which we call a *Regularization Network* (RN).

A special case of regularization networks are the Radial Basis Functions (RBFs) networks, where the function $G(\mathbf{x})$ has radial symmetry.

An interesting and physiologically plausible set of basis functions is a set of multidimensional Gaussians where

$$h(\|\mathbf{x} - \mathbf{x}_i\|) \equiv e^{-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2}} \tag{6.6}$$

Equation ?? can then be written as

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}\right) + p(\mathbf{x}) \tag{6.7}$$

where $\|\mathbf{x} - \mathbf{x}_i\|^2$ is the distance of the input vector $\mathbf{x}$ from its center $\mathbf{x}_i$. Since the exponential is separable, any multidimensional Gaussian can be substituted by the product of one-dimensional Gaussians, and any multidimensional center can be factorized in terms of one-dimensional centers. For example, a $2D$ Gaussian function centered on $\mathbf{x}_i$ can be written as:

$$h(\|\mathbf{x} - \mathbf{x}_i\|) \equiv \exp(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}) = \exp(-\frac{(x^{(1)} - x_i^{(1)})^2}{2\sigma_1^2}) \exp(-\frac{(x^{(2)} - x_i^{(2)})^2}{2\sigma_2^2}) \tag{6.8}$$

where $\mathbf{x} \equiv (x^{(1)}, x^{(2)})$. Thus an RBF network with Gaussian basis functions can be conceived of as a set of hierarchically organized "hidden-layer" units, with bell-shaped *receptive fields*. Units at different layers have connections, the strength of which is determined during the training phase.

In the version of RBF described so far the network has one hidden unit (basis function) for each data point. It has been argued by a number of authors (Moody and Darken, 1989; Poggio and Girosi, 1990a) that the number of hidden units can be drastically reduced, and good approximation can be achieved with a number of basis functions which is smaller than the number of data points, especially in the case in which the data set contains clusters of data points very close to each other. Therefore the following extension of the standard RBF technique has been proposed:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i h(\|\mathbf{x} - \mathbf{t}_i\|) + p(\mathbf{x}) \tag{6.9}$$

where $\{\mathbf{t}_i\}_{i=1}^n$ is a new set of free parameters, called *centers*, that can be either fixed by the user according to some heuristic or prior knowledge, or learned via gradient descent together with the coefficients c_i. The centers now play the role of "prototypes" against which a new input is matched in order to be classified. Notice that the distance between an input and a center is the standard euclidean distance, which assigns equal weight to all the variables. In many cases, this is not a good choice, and a more flexible model has been proposed (Poggio and Girosi, 1990a), that takes in account the fact that certain features can be more relevant than others. The new model has the form:

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i h(\|(\mathbf{x} - \mathbf{x}_i)\|_W) + p(\mathbf{x}) \tag{6.10}$$

where W is a (possibly rectangular) matrix of free parameters, and we have defined the *weighted norm* as

$$\|(\mathbf{x} - \mathbf{x}_i)\|_W^2 = (\mathbf{x} - \mathbf{x}_i)^T W^T W (\mathbf{x} - \mathbf{x}_i) \tag{6.11}$$

Finding the optimal weights W for this norm is equivalent to transforming appropriately the input coordinates, and this corresponds to task-dependent dimensionality reduction. In other words, the weights of the norm determine the significance of the visual features that comprise the input to the recognition system.

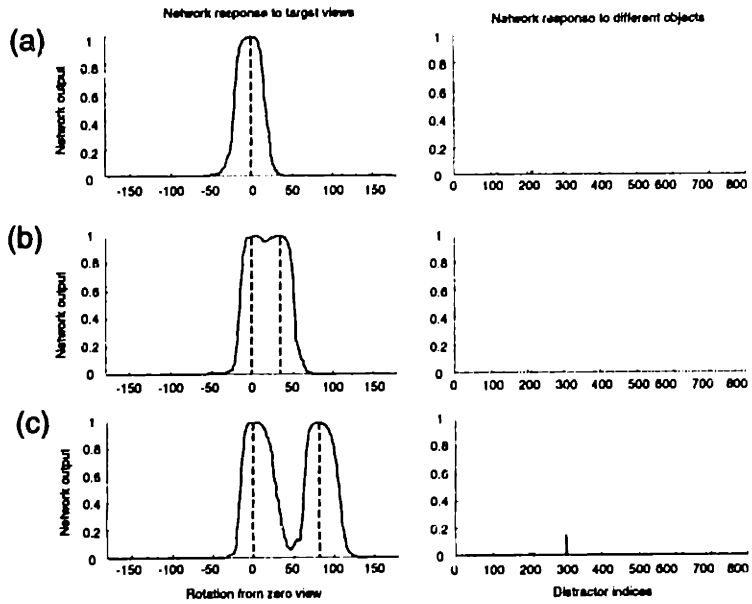

Figure 6-2: Performance of an RBF net trained with different number of views using as input x, y coordinates of wire-like objects. The left column shows the network output for different views of the learned object, the right column the output for different objects. (a) Network trained with a single view (b) Network trained with 2 views 40° apart (c) Network trained with 2 views 80° apart

6.1.3 Visual Features For Recognition

The RBF model proposed by Poggio and Edelman allows us to categorize correctly an object by comparing its representation to those of a few "views" stored in memory (equivalent to the centers of the RBF functions). The model does not specify what the nature of the stored internal representation should be. The only requirement is that the representation be detailed enough to code the important characteristics of the object. In fact, as preliminary simulations have shown, x, y coordinates and segment orientations could be used as recognition features (Poggio and Edelman, 1990; Logothetis et al., 1994). The performance of an RBF-type network, when x, y coordinates are used for wire-like objects, is given in Figure 6-2. The network

performs satisfactorily, however this does not shed any light on the plausibility of these parameters as features used by the visual system.

As pointed out by Marr, the choice of the internal representation is extremely important, since it makes some of the information in the original stimulus more explicit and immediately available (Marr, 1982). An optimal representation should code just enough detail to be able to distinguish between two similar objects, while discarding the information common to all the objects, and therefore not useful in the recognition task.

Depending on the specifics of the recognition task, and on the class of objects involved, we can hypothesize different possible features or "characteristic regions" that could be the ones used in the representation at the higher level. In the case of the basic level categorization, volumetric primitives could be used for describing most of the objects we encounter in our daily lives (Cooper et al., 1992a). On the other hand, if the task is a subordinate (within-class) level classification, where the different objects all have the same components, what probably needs to be coded is a combination of distances between characteristic features. This means that the representation could code the underlying metric of the object. But even both these sets of features used together would not guarantee the ability of efficiently coding objects for which only holistic, configurational shape is available, as could be the case of the amoeba-like objects.

Most likely a combination of these and other types of features is used in the representation. We would like to determine which is the representation used by the human visual system, since its performance in recognition exceeds by far any of the artificial systems. In order to simplify our task, the "feature-search" was specific for a limited class of objects. Wire-like objects and amoebas are used in subordinate level classification experiments. The former are obtained by the juxtaposition of specific volumetric primitives, while the latter are not easily decomposed into subelements.

Unlike other computational approaches, we are not looking for any set of features that allowed us to perform the recognition task, rather we want to find biologically plausible features that can also be used to generate filters for artificial recognition

systems. The output of such filters can be used, for example, as an input to the RBF-type networks. In order to achieve this goal, neural responses from units in IT which show differential activity modulation when presented with different objects are analyzed and classifyed.

6.2 Physiological experiments that generate the database for the current analysis

Since we intend to look for "biologically plausible" features, we want to relate our findings to the responses of inferotemporal (IT) cortex cells. We have chosen this area because all the evidence seems to indicate it as the most likely area where object representation is stored. We will utilize a large database consisting of 1100 cells recorded in area IT, at the laboratory of Nikos Logothetis at Baylor College of Medicine. A more detailed description of the physiological data can be found in Chapter II, and in Logothetis and Pauls (1995), Logothetis et al. (1995), Pauls et al. (1996).

The data set was collected from 2 awake, behaving monkeys (*Macaca Mulatta*). The cells activity was collected while the monkeys were performing a recognition task in which two phases could be distinguished. In the first phase, an object was presented on the screen while undergoing a motion sequence for $\sim$ 500ms. This motion sequence of two-dimensional views conveyed an impression of 3D shape through structure from motion. During this time, the animal was trained to memorize the object, subsequently referred to as the target. After a 800ms delay, the following phase started. A number of objects were shown statically, one after the other. At the end of each presentation, the monkey had to press one of two levers, depending on whether the object was the target, or one of many distractors. The test object had to be classified as a target, even if it was presented from a viewpoint different from the ones shown in the training sequence [2].

[2]For more detail on the protocol see Chapter II

A large number of the neurons recorded during the testing phase of the recognition task was responsive to visual stimuli. Neurons mainly increased their activity when presented with the images, and only few were inhibited by the stimulus presentation. The stimuli used to plot the neurons were mostly belonging to the wire-like or amoeba-like objects classes on which the monkeys had learned to perform the recognition task.

Of particular interest to our analysis are two different categories of neurons that happened to be stimulus selective. The first category includes neurons that responded only to a subset of views of a particular target, while being insensitive to the presentation of any other distractors. The second category of neurons did not respond to a single object, but rather to a subset of the distractors used in the presentation. Sometimes the activity in this latter class was also elicited by some of the target views. Both these cases arouse our interest because they are likely to be coding some characteristics of the object (or objects, if more than one) to which they respond. The pattern of activity on the retina changes greatly for the different excitatory stimuli, but the neurons seem to be able to extract some common information while eliminating irrelevant data.

6.3 A Recognition Model Based on Nonlinear Interpolation Between Stored 2D views

Recognition of specific objects, such as recognition of a particular face, can be based on representations that are object-centered, such as 3D structural models. Alternatively, a 3D object may be represented for the purpose of recognition in terms of a set of views. This latter class of models is biologically attractive because model acquisition – the learning phase – is simpler and more natural.

A simple model for this strategy of object recognition was proposed by Poggio and Edelman (Poggio and Edelman, 1990). They showed that, with few views of an object used as training examples, a classification network such as a Gaussian Radial Basis Function network can learn to recognize novel views of that object. This is particularly marked in the case of views obtained by in-depth rotation of

the object (translation, rotation in the image plane, and scale, are probably taken care by object-independent mechanisms). The model, sketched in Figure 6-1, makes several predictions about limited generalization from a single training view, and about characteristic generalization patterns between two or more training views (Bülthoff and Edelman, 1992). Psychophysical and neurophysiological results support the main features and predictions of this simple model. For instance, in the case of novel objects, it has been shown that when subjects --both humans and monkeys- are asked to learn an object from a single unoccluded view, their performance decays as they are tested on views farther away from the learned one (Bülthoff and Edelman, 1992; Tarr and Pinker, 1991; Logothetis et al., 1994). Additional work has shown that even when 3D information is provided during training and testing, subjects recognize in a view-dependent way, and cannot generalize beyond 40 degrees from a single training view (Sinha and Poggio, 1994).

Even more significantly, recent recordings in inferotemporal cortex (IT) of monkeys performing a similar recognition task with wire-like and amoeba-like objects, revealed cells tuned to specific views of the learned object (Logothetis et al., 1995). The tuning, an example of which is shown in Figure 6-8, is presumably acquired as an effect of the training to views of the particular object. Thus an object can be thought as represented by a set of cells tuned to several of its views, which is consistent with the findings of others (Wachsmuth et al., 1994). This simple model can be extended to deal with symmetric objects (Poggio and Vetter, 1992) as well as objects which are members of a *nice*[3] class: in both cases, generalization from a single view may be significantly greater than for objects such as the wire-like objects used in the psychophysical and physiological experiments (Vetter et al., 1995).

The original model of Poggio and Edelman has a major limitation: it does not specify which features are inputs to the view-tuned units, and what is stored as a representation of a view in each unit. The simulation data they presented employed features such as the x, y coordinates of the object vertices in the image plane, or the angles between successive segments. This representation, however, is biologically

[3]as defined in Vetter el al. (1995)

implausible, and particularly so for objects that have easily detectable vertices and measurable angles, like wire-like. Here we suggest a view representation which is more biologically plausible, and which applies to a wider variety of cases. We will also show that this extension of the Poggio-Edelman model leads to properties consistent with the cell response to the same objects.

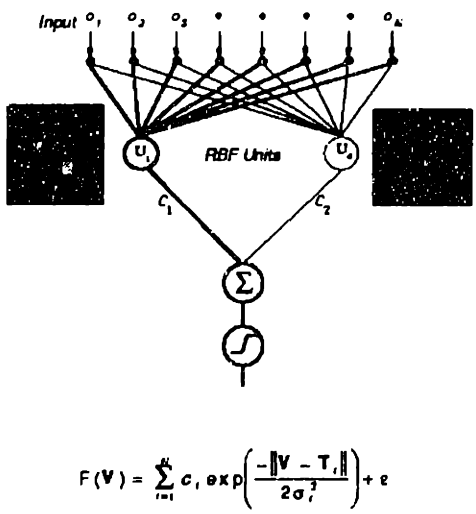

Figure 6-3: Radial Basis Function network

6.3.1 Whole image prototypes

Figure 6-3 shows a schematic of the system in its simpler form. It is equivalent to a one layer network of Radial Basis Functions (RBF) units with a thresholded output. The input to the network is a 64x64 array $\mathbf{V}$, the grey-level image obtained by viewing a given object from a specific view[4]. Positive examples (views of the objects to be learned, also called *targets*, are required to have output 1 and negative examples (views of any other object also called distractors) are required to have output 0. The

[4]Wire-like and amoeba-like objects viewed from different viewpoints were rendered as greyscale images. Care was taken so that all the objects were centered in the image and were approximately of the same scale, a preprocessing stage that takes care of some image transformations (see Discussion). The images were then subsampled so to obtain a 64x64 array. These reduced images were used to train and test the recognition system described below. Figure 6-3 shows an example of the images used: the two images belong to the same object rotated around the vertical axis of 60°.

whole image of the object is used as the prototype T_i. The number of prototypes, or centers, is determined by the number of positive examples used in training. Figure 6-3 shows a case where two centers, T_1 and T_2, are used for the given target. The centers represent views of the object to be learned, separated by 60° along the horizontal axis. For each image $\mathbf{V}$, presented as input to the network, the euclidean distance between the image in consideration and the image used as prototype (i.e. as center of the RBF unit) is computed. A gaussian function weights these values with appropriate sigma (chosen to be of the order of the mean value of the distances). These weighted distances from one or more prototypes (in the case of views of the same object) are used to train a classification network. The output weights c_i and threshold are learned in this process. The output of the network before the threshold can be expressed as (see also equation 6.10).

$$F(\mathbf{V}) = \sum_{i=1}^{N} c_i \exp\left(\frac{-\|\mathbf{V} - \mathbf{T}_i\|^2}{2\sigma_i^2}\right) \tag{6.12}$$

In the results presented here, one or two positive examples $N = 1, 2$ are presented in the training phase.

In order to obtain a more robust representation in a second set of simulations, we added a layer of units to the network, comprised of a set of four first-order oriented filters spaced 45° apart. For each view of the target, the network now had four centers one for each of the filtered images. The remaining procedure was unchanged (see Figure 6-4b).

We analyzed the generalization capabilities of the two versions of the network, with and without the layer of filters, by presenting to it equally spaced views of the object used as targets, together with views of 200 other different objects belonging to the same family. The results are shown in Figure 6-4a, in the case of training on raw images and in Figure 6-4b, in the case in which instead of raw images for each presented positive example we used four centers (each corresponding to the training view filtered through a different filter).

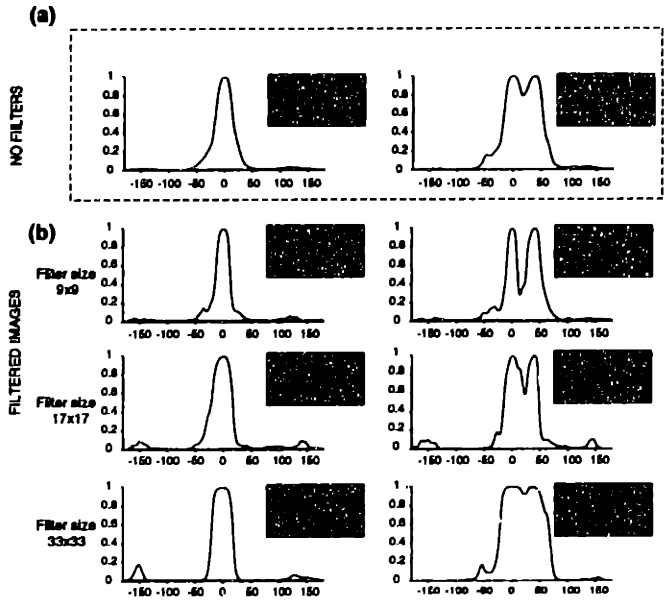

Figure 6-4: Generalization behavior of the network. The network output is plotted against view of the object. Dashed vertical lines represent the views used in the training set. In the small inset we see the response to the 200 negative examples tested. (a) No filters used. (b) Preprocessing of images with filters of different sizes.

6.4 A model of view-tuned units

We now propose a model for these view-tuned units. Our approach consists of representing a view in terms of a few local features, which can be regarded as local configurations of grey-levels. Choose one point in the image of the object: a feature vector is computed by filtering the image with a set of filters, with small support, centered at the chosen location. The vector of filter responses is used as a description of the local pattern in the image. Four such points are chosen, for example, in the image of Figure 6-8a, where the white squares indicate the support of the bank of filters which are used. Since the support is local but finite, the value of each filter depends on the pattern contained in the support, and not only on the center pixel; since there are several filters, one expects that the vector of values may uniquely represent the local feature, for instance a corner of the wire-like object.

We use filters that are somehow similar to oriented receptive fields in V1 (though it is far from being clear whether some V1 cells behave as linear filters). The ten

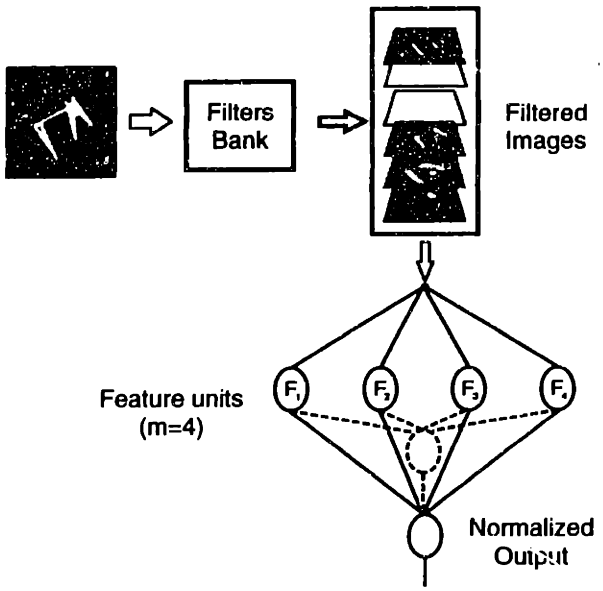

Figure 6-5: Model overview: during the training phase the images are first filtered through a bank of steerable filters. Then a number of image locations are chosen by an attentional mechanism and the vector of filtered values at these locations is stored in the feature units.

filters used are the same steerable filters (Freeman and Adelson, 1991) suggested by Ballard and Rao (Rao and Ballard, 1995; Leung et al., 1995). The filters are chosen to be a basis of steerable filters up to the third order, including the zero order Gaussian. The filters, shown in Figure 6-6, are defined as:

$$G_n^{\theta_n} = \frac{\partial^n}{\partial x^n} G(x,y) \quad \begin{array}{l} n = 0,1,2,3 \\ \theta_n = 0,...,k\pi/(n+1), k = 1,...n \end{array} \tag{6.13}$$

where

$$G(x,y) = \exp^{-(x^2+y^2)} \tag{6.14}$$

Therefore for each one of the chosen locations m in the image we have a 10-value array $\mathbf{T}_m$ given by the output of the filters bank.

$$\mathbf{T}_m = ((I * G_0^0)|_m, (I * G_1^{\pi/2})|_m, ..., (I * G_3^{3\pi/4})|_m) \tag{6.15}$$

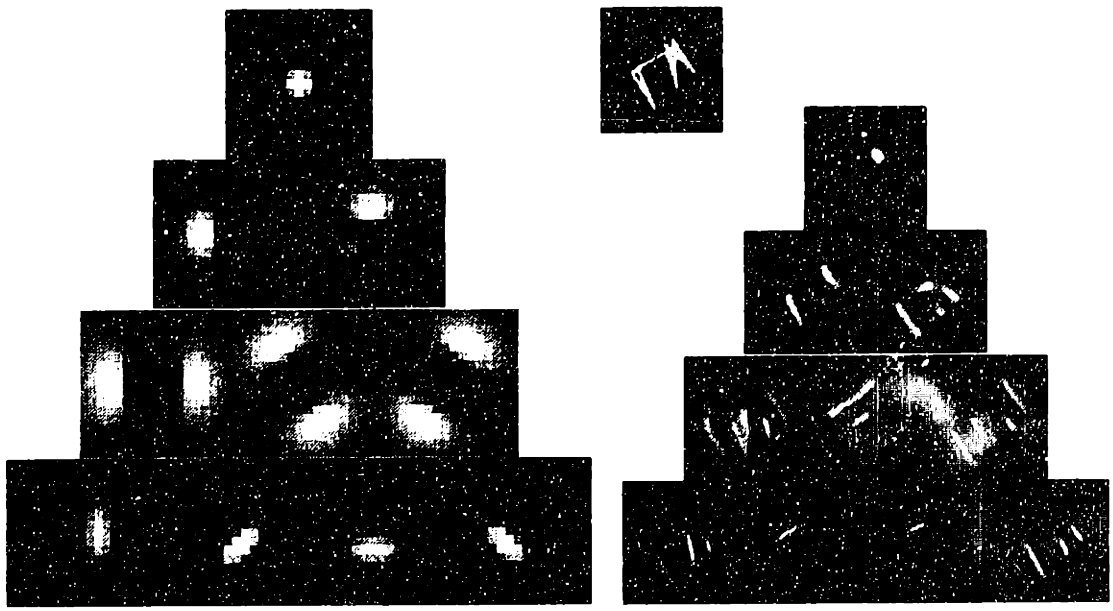

Figure 6-6: The ten oriented filters used are shown on the left. The result of filtering one of the images used with each of the filters on the right.

The representation of a given view of an object is then built as follows: First $m = 1, ..., M$ locations are chosen. For each of these M locations, the 10-valued vectors $\mathbf{T}_m$ are computed and stored. These M vectors, with M typically between 1 and 4, form the representation of the view which is learned and commited to memory.

How are the locations chosen? Precise location is not critical, and it turns out that the feature locations can be chosen almost randomly. Of course, each specific choice will influence properties of the unit, but precise location does not affect the qualitative properties of the model, as verified in simulation experiments. Intuitively, features should be centered at salient locations in the object, where there are large changes in contrast and curvature. We have implemented (Riesenhuber and Bricolo, in preparation) a simple attentional mechanism that chooses locations close to edges with various orientations[5]. The locations shown in Figures 6-8 and 6-9 are obtained using this unsupervised technique. We emphasize however that none the results and conclusions here reported depend on the specific location of the feature or the precise

[5]A saliency map is at first constructed as the average of the convolutions of the image with four directional filters (first order steerable filters with $\theta = 0, ..., k\pi/(4), k = 1, .., 4$). The locations with higher saliency are extracted one at the time. After each selection, a region around the selected position is inhibited to avoid selecting the same feature over again.

procedure used to choose them.

Thus far, we have described the learning phase, and how views are represented and stored. When a new image $\mathbf{V}$ is presented, recognition takes place in the following way. First, the new image is filtered through the bank of filters. Thus at each pixel location i, we have the vector of values $\mathbf{f}_i$ provided by the filters. Now, consider the first stored vector $\mathbf{T}_1$. The closest $\mathbf{f}_i^*$ is found searching over all i locations, and the distance $D_1 = \|\mathbf{T}_1 - \mathbf{f}_i^*\|$ is computed. This process is repeated for the other feature vectors $\mathbf{T}_m$ for $m = 2, ..., M$. Thus, for the new image $\mathbf{V}$, M distances D_m are computed; the distance D_m is therefore the distance to the stored feature $\mathbf{T}_m$ of the closest image vector searched over the whole image.

The model uses these M distances as exponents in M Gaussian units. The output of the system is a weighted average of the output of these units with an sigmoidal output non linearity:

$$\mathbf{Y_V} = h\left(\sum_{m=1}^{M} c_m e^{-\frac{D_m^2}{2\sigma^2}}\right) \tag{6.16}$$

In the simulations presented here, we estimated σ from the distribution of distances over several images; the c_m are $c_m = M^{-1}$, since we have only one training view; h is $h(x) = 1/(1 - e^{-x})$.

The distance for both the target and the distractors singularly for each feature is shown in Figure 6-7b. Notice that if only a single prototype is used, the output varies a lot according to the specificity of the prototype (some being less selective than others), and also accordingly to the distractor set. In Figure 6-7c, we see the result obtained by simply averaging the output of the four feature detectors shown before.

In Figure 6-8b, we see the results obtained by a simple linear combination of the output of the four feature detectors, followed by the sigmoidal non linearity, in the case of a different object. We have also experimented with a multiplicative combination of the output of the feature units. In this case, the system performs an AND of the M features. If the response to the distractors is used to set a threshold for classification,

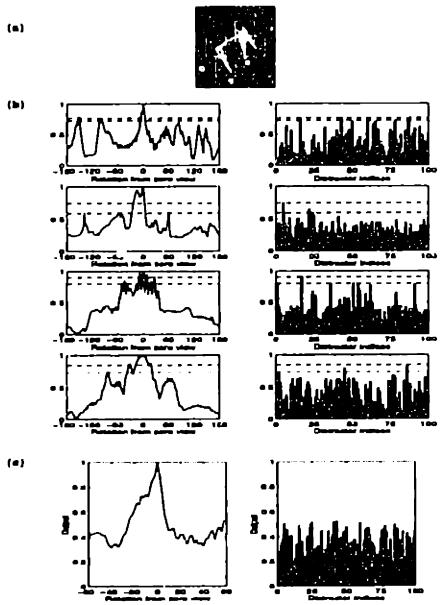

Figure 6-7: (a) Target image with the four subimages used as features highlighted (b) In the left column the network output is plotted against view of the object, in the right column the response to the 200 negative example. (c) Response of the network to the combined response of all the features for target and distractors

then the two versions of the system behave in a similar way. Similar results (not shown) are obtained using other kinds of objects.

6.5 Results

6.5.1 Comparison between view-tuned units and cortical neurons

Electrophysiological investigations in alert monkeys, trained to recognize wire-like objects presented from any view, show that the discharge rate of many IT neurons is a bell-shaped function of orientation, centered on a preferred view (Logothetis et al., 1995). The properties of the units described here are comparable to those of the cortical neurons (see Figure 6-8). The model was tested with exactly the same objects used in the physiological experiments. As a training view for the model, we used the view preferred by the cell (the cell became tuned presumably as an effect of

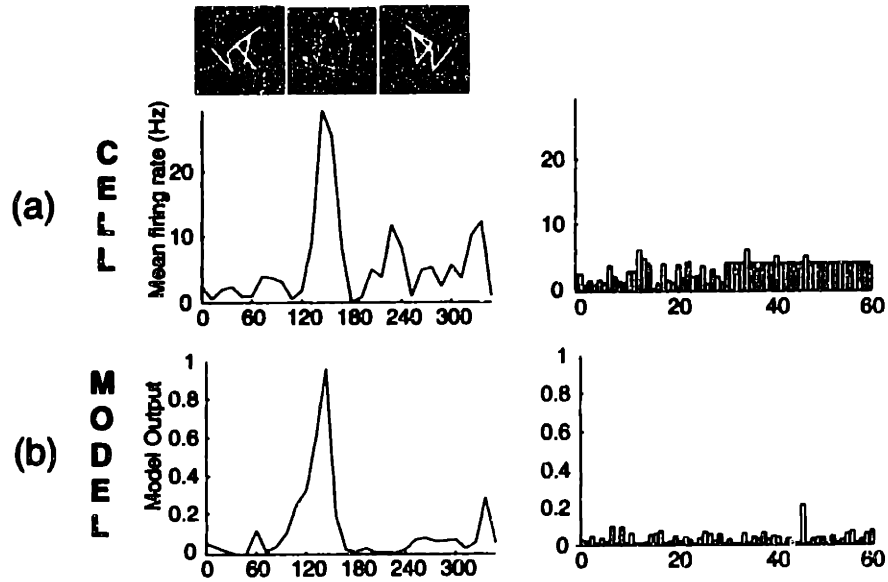

Figure 6-8: Comparison between a model view-tuned unit and cortical neuron tuned to a view of the same object. (a) Mean spike rate of an inferotemporal cortex cell recorded in response to views of a specific wire-like object and to a set of distractor objects (Logothetis and Pauls, personal communication). (b) Model response for the same set of objects. This is representative for other cells we have simulated, though there is considerable variability in the cells (and the model) tuning.

training during which the monkey was shown several views of this object).

6.5.2 Occlusion experiments

What physiological experiments could lend additional support to our model? A natural question concerns the behavior of the cells when various parts of the object are occluded. The predictions of our model for a specific object and a specific choice of feature units ($m = 4$) and locations in this particular case are given in Figure 6-9.

Simulations show that the behavior depends on the position of key features with respect to the occluder itself. Occluding a part of the object can drastically reduce the response to that specific view (Figure 6-9b(ii-iv)) because of interference with more than one feature. But since the occluded region does not completely overlap with the occluded features (considering the support of the filters), the presentation of this region alone does not always evoke a significant response (Figure 6-9b(iii-v)).

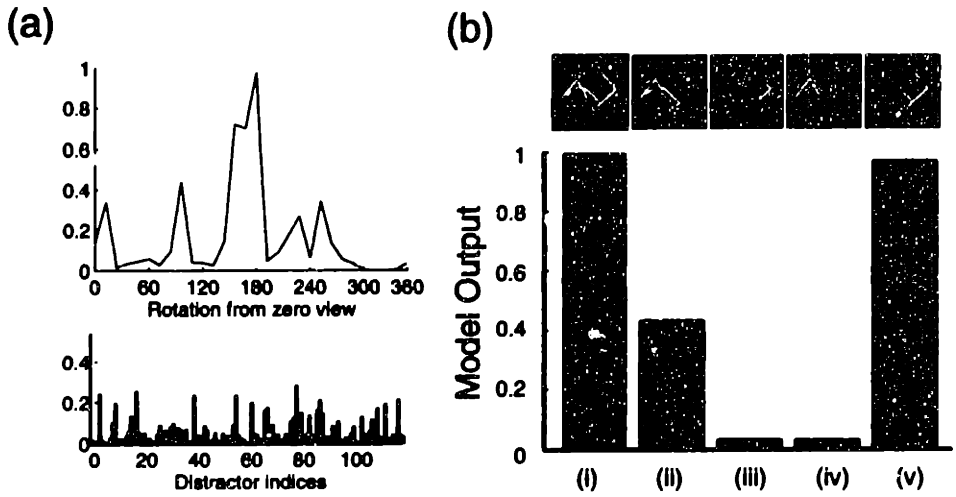

Figure 6-9: (a) Model behavior in response to a learned object in full view (highlighted on the learned image are the positions of the four features) at different rotations and to various different objects (distractors), (b) Response dependence on occluder characteristics: (i) object in full view at learned location, (ii) object occluded with a small occluder, (iii) occluded region in (ii) presented in isolation, (iv-v) same as (ii-iii) but with a larger occluder.

6.6 Discussion

The Poggio and Edelman model was designed specifically for wire-like objects, and did not explicitly specify how to compute the response for any object and image. Here, we fill this gap and propose a model of these IT cells that become view-tuned as an effect of training. The key aspect of the model is that it relies on a few local features (1-4) that are computed and stored during the training phase. Each feature is represented as the set of responses of oriented filters at one location in the image. During recognition, the system computes a robust conjunction of the best matches to the stored features.

Clearly, the version of the model described here does not exploit information about the geometric configuration of the features. This information is available once the features are detected, and can be critical to performing more robust recognition. We have devised a model of how to use the relative position of the features $\mathbf{f}_i^*$ in the image. A biologically plausible translation and scale invariant model can be made

by using a network of cells with linear receptive fields, which is similar in spirit to a model proposed for spatial representation in the parietal cortex (Pouget and Sejnowski, 1996). Interestingly enough, this additional information is not needed to account for the selectivity and the generalization properties of the IT cells which we have considered so far. The implication is that IT cells may not be sensitive to the overall configuration of the stimulus, but rather to the presence of moderately complex local features (according to our simulations, the number of necessary local features is greater than one for the most selective neurons, such as the one of Figure 6-8a). Scrambling the image of the object should therefore preserve the selectivity of the neurons, *provided* this can be done without affecting the filtering stage. In practice, this may be very difficult. Though ours is still an incomplete neuronal model, it is already being used to make useful predictions for physiological experiments which are currently underway.

Chapter 7

Conclusions

In summary, this thesis addresses the issue of *characterizing the representation of simple, isolated, rigid objects for shape-based recognition.* In addressing this issue, we make an effort to merge information from various fields of vision research — such as psychophysics, neurophysiology, and computational modeling. The main results can be summarized as follows

- **Psychophysical results**
  - In Experiment 1, we tested recognition performance for a number of view of objects against distractors with varying degree of similarity. At low noise level, i.e. when objects are very similar, subjects are unable to discriminate targets from distractors at all tested viewpoints. This result suggests that the view tuning is obtained by storing a prototype based on the image. As the similarity between targets and distractors decreases the discrimination becomes easier and is independent of the amount of distortion in the image plane between the original center-view and the object in consideration, whether a target or a distractor. This confirms that the frame of reference used in subordinate level classification is viewpoint-dependent.
  - In subsequent experiments (Experiment 2 to Experiment 5), we tested recognition of a single view of an object in more detail. The results suggest that, at least for the class of objects used in our experiments, this

representation may be strongly configurational, weighting some characteristic parts of the image more than others.

- A final set of experiments tested the dependence of object representation on object-dependent transformation. We showed that the representation is indeed size- and position dependent even if the effect is not as robust as in the case of object specific transformations.

- **Monkey Physiology**

  - Even when complete information about the structure of an object is available to the subject, recognition at the subordinate level depends on the object's attitude, both for monkeys as well as for human subjects.

  - A memory-based, viewer-centered recognition system is not an implausible mechanism for object-constancy. Both theoretical work, and the results presented here, suggest that only a small number of object views need to be stored in order to achieve perceptual invariance.

  - A small population of IT neurons has been found to respond selectively to individual members of the object-classes tested in this study. The response of some neurons is a function of the object's view. The discharge rate of many IT neurons is found to be a bell-shaped function of orientation centered on a preferred view and view-tuning is observed only for those views that the monkeys can recognize.

  - A subset of the view-selective units were also tested for position and size changes of the preferred view. In our experiments, the responses of IT neurons lie on a continuum from cells invariant to either scale, position, or both, to cells with varying degrees of sensitivity to these parameters.

- **Computational model**

  - We propose a model of these view-tuned units that is consistent with physiological data from single cell responses. Our approach consists of repre-

senting a view in terms of a few local features, computed by V1-like cells, which can be regarded as local configurations of grey-levels.

- The model units are view-selective. Their behavior is comparable to that of view-tuned cells as characterized by their response to rotation in depth of the objects to which they are tuned and by the relative response to distractors.
- The model also predicts the complex behavior seen for partly occluded objects. An occlusion may or may not disrupt the response, depending on the number and position of the extracted features.

Appendix A

Construction of basic set of stimuli

A.1 Wire-like objects

All stimuli resembled paper clips unbent to make a meaningless sculpture and were rendered, when not otherwise specified in the specific experiment, as orthographic projection of shaded 3D objects using the "GL Library" on a SGI Indigo2. The stimuli were a set of non self occluding objects constructed from cylindrical components serially connected end-to-end. These objects are similar to those used by Bülthoff and Edelman (Bülthoff and Edelman, 1992). Two main classes of objects were defined.

Objects belonging to *Class I* were composed of cylindrical segments serially connected so to satisfy the following constraints:

1. Each object was composed of 7 segments
2. The segments did not intersect
3. Each connection angle between two successive segments was chosen pseudo-randomly between 30° and 150°.
4. The three moments of inertia of each object were within 10% of their average value.
5. The segments were all of the same length

Objects belonging to *Class II* were derived from objects belonging to Class I by adding varying amounts of noise to the vertices. Therefore, the objects belonging to Class II did not satisfy constraints 3 to 5.

Figure A-1: Examples of stimuli and mask used in both psychophysical and physiological studies.

An unlimited number of novel objects with controlled complexity could be generated in any of the two classes. The constraint on the moment of inertia forced the object not to have a principal axis. The frontal view of the object (also called in this thesis 0° view) was arbitrarily selected. Different sets of objects were generated for each experiment. There were no requirements on symmetry (Figure A-1).

When not otherwise specified in the single experiment, the color of the object was gray. Ambient white light was simulated to illuminate the surface. The default direction of the light originated in the upper left corner.

The objects generated as members of this class were used in all the psychophysical and physiological experiments presented in this thesis, as well as for the computer simulations.

Appendix B

Principal Component Analysis for Raw Image Values

The goal of the Karhunen-Loewe expansion (or Principal Components Analysis) is to reduce the dimensionality of data sets with large number of highly correlated independent variables. The original set of variables is linearly transformed into a substantially smaller set of uncorrelated variables without loosing the original information. This new variables form a set of orthonormal basis function that still span the original space. This new set of basis can be ordered so that the first represents the direction along which the original data varies most. The first few basis of this ordered set could be used to drastically reduce the dimensionality of the data while they still account for most of its variance. They are called the *principal components.*

Here we apply the technique to image analysis where the original variables (the pixels) are highly correlated. This approach enables us to express a set of m input images $I_1, I_2, ..., I_m$ as a linear combination of a set of characteristic feature images $e_1, e_2, ..., e_m$ called eigenimages . Each of this eigenimages can be considered as one of the axes of the space spanned by the initial set. Each vector e_k has a parameter value associated with it, the eigenvalue, that specifies how much of the variance of the original set is accounted for by expressing the original images in terms of each single eigenvector. The highest the eigenvalues, the higher the variability is taken into account. To reduce the dimensionality of the set only few of the eigenimages

(the one with highest eigenvalues) are used to express the data.

The methods treats images as general 1-dimensional arrays with no assumptions about configurational properties of objects represented.

The paradigm can be divided up in 3 steps:

1. Compute correlation matrix C for the original images.

2. Compute eigenvalues and eigenvectors of the correlation matrix C .

3. use the eigenvectors to compute the eigenimages.

Let $I_1, I_2, ..., I_m$ be our set of m original images represented as vectors of dimensions $1xN^2$. The average image M of the set is defined as

$$M = 1/m * \sum_{i=1}^{m} I_i$$

Consider the vectors obtained by subtracting each image from the mean.

$$H_i = I_i - M \tag{B.1}$$

We want to obtain a set of vectors that best describe the distribution of the data in space. The k-th vector is chosen such that

$$\lambda_k = 1/m \sum_{i=1}^{m} (e_k^T H_i)^2 \tag{B.2}$$

is a maximum subject to

$$e_l^T e_k = \delta_{lk} = \begin{cases} 1 & \text{if l=k} \\ 0 & \text{otherwise} \end{cases} \tag{B.3}$$

That is the vectors e_k and the scalars λ_k are the eigenvalues and the eigenvectors of the covariance matrix

$$C = 1/m \sum_{i=1}^{m} H_i H_i^T = HH^T \tag{B.4}$$

The matrix C is N^2xN^2 with N^2 eigenvectors and corresponding eigenvalues. But if the dimension of the space spanned by C is less than the actual dimension of the matrix ($m < N^2$) the actual number of meaningful eigenvectors (with eigenvalues different from 0) is actually N. In this case we can find the m eigenvectors we are interested in by solving a smaller system.

Let's consider the matrix $K = H^T H$ of dimensions mxm. It has m eigenvectors v_i with corresponding eigenvalues μ_i that satisfy the following:

$$Kv_i = \mu_i v_i$$

$$H^T H v_i = \mu_i v_i$$

This is a much smaller problem to solve if $m << N^2$. Now if we premultiply both sides by H, we obtain:

$$HH^T H v_i = \mu_i H v_i$$

That is

$$CHv_i = \mu_i Hv_i$$

That is the eigenvectors of the image correlation matrix (eigenimages) are given as linear combinations of the images themselves using the eigenvector of matrix H as coefficients.

The m meaningful eigenvectors of the matrix C are $e_i = Hv_i$ and the corresponding eigenvalues are $\lambda_i = \mu_i$. The eigenimages are therefore obtained by

$$e_k = \sum_{i=1}^{m} v_{ik} h_k \tag{B.5}$$

Examples

We have performed the PCA decomposition on a set of 120 images of wire-like objects that have been used as stimuli in the physiological experiments described above. The input set had dimensions $N = 4096$ but spanned a $120th$ dimensional

space. We computed the corresponding 120 eigenimages. The eight eigenimages with the highest eigenvalues are shown in Figure B-1. We then projected the $120th$ original images on the space spanned by the first 10 eigenimages obtaining a set of points in a 10 dimensional space.

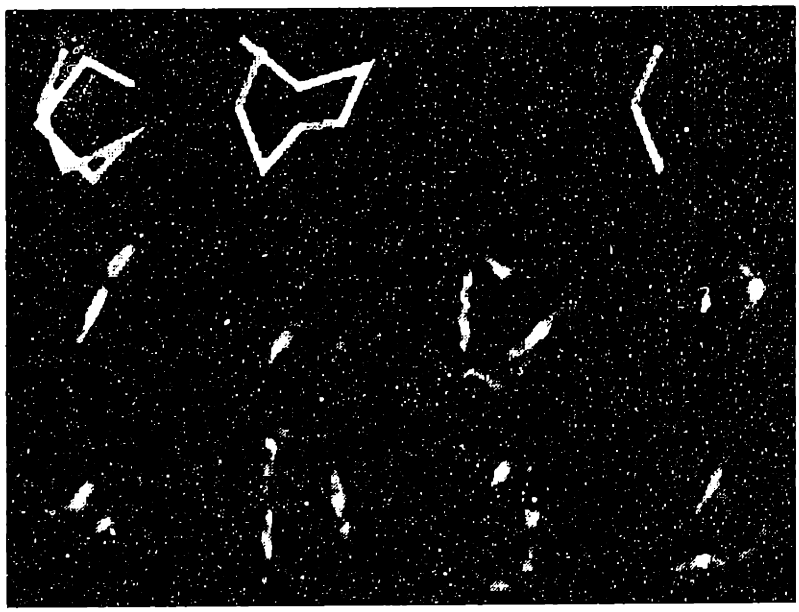

Figure B-1: Shown in the first row two of the original images and the mean image of the whole set. In the two subsequent rows from left to right are displayed the eight eigenimages with the eight highest eigenvalues.

Bibliography

Albright, T. and Gross, C. (1990). Do inferior temporal cortex neurons encode shape by acting as fourier descriptor filters? *Proceedings of the International Conference on Fuzzy Logic and Neural Networks*, pages 375–378.

Besner, D. (1983). Visual pattern recognition: Size preprocessing re-examined. *Quarterly Journal of Experimental Psychology A*, 35A:209–216.

Besner, D. and Coltheart, M. (1975). Same-different judgments with words and nonwords: The differential effects of relative size. *Memory & Cognition*, 3(6):673–677.

Besner, D. and Coltheart, M. (1976). Mental size scaling examined. *Memory & Cognition*, 4(5):525–531.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147.

Biederman, I. and Cooper, E. (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20:585–593.

Biederman, I. and Cooper, E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):121–133.

Bricolo, E., Pauls, J., and N.K., L. (1995). The role of inferior temporal cortex in visual object recognition. In Herrmann, H., Wolf, D., and Poppel, E., editors,

Workshop of Supercomputing and Brain Research, pages 225–241. World Scientific.

Bruce, C., Desimone, R., and Gross, C. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology*, 46:369–384.

Bülthoff, H. and Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Science. USA*, 89:60–64.

Bülthoff, H., Edelman, S., and Tarr, M. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, 3:247–260.

Bundsen, C. and Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3):214–220.

Carman, G., Drury, H., and VanEssen, D. (1995). Computational methods for reconstructing and unfolding the cerebral cortex. *Cerebral Cortex*, 5(6):506–517.

Cavanagh, P. (1978). Size and position invariance in the visual system. *Perception*, 7:167–177.

Cohen, L., Gray, F., Meyrignac, C., S., D., and Degos, J.-D. (1994). Selective deficit of visual size perception: Two cases of hemimicropsia. *Journal of Neurological and Neurosurgical Psychiatry*, 57:73–78.

Cooper, E. E., Biederman, I., and Hummel, J. E. (1992a). Metric invariance in object recognition: A review and further evidence. *Canadian Journal of Psychology*, 46(2):191–214.

Cooper, L. and Schacter, D. (1992). Dissociations between structural and episodic representations of visual objects. *Current Directions in Psychological Science*, 1(5):141–146.

Cooper, L., Schacter, D., Ballesteros, S., and Moore, C. (1992b). Priming and recognition of transformed three-dimensional objects: effects of size and reflection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(1):43–57.

Corballis, M. (1988). Recognition of disoriented shapes. *Psychological Review*, 95(1):115–123.

Cormack, R. (1985). The computation of retinal disparity. *Perception and Psychophysics*, 37:176–178.

Cowey, A. and Gross, C. (1970). Effects of foveal prestriate and inferotemporal lesions on visual discrimination. *Experimental Brain Research*, 11:128–144.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation. *Numer. Math*, 31:377–403.

Damasio, A. (1990). Category-related recognition defects as a clue to the neural substrates of knowledge. *Trends in Neuroscience*, 13:95–99.

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

Desimone, R., Albright, T., Gross, C., and Bruce, D. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4:2051–2062.

Dill, M., Wolf, R., and Heisenberg, M. (1993). Visual pattern recognition in drosphila involves retinotopic matching. *Nature*, 365:751–753.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley.

Edelman, S. and Bülthoff, H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, 32(12):2385–2400.

Edelman, S. and Poggio, T. (1992). Bringing the grandmother back into the picture: A memory-based view of object recognition. *International journal of Pattern Recognition and Artificial Intelligence*, 6(1):37–62.

Fiser, J. and Biederman, I. (1995). Size invariance in visual object priming of gray scale images. *Perception*, 24(7):741–748.

Foster, D. H. and Kahn, J. I. (1985). Internal representations and operations in the visual comparison on transformed patterns: Effects of pattern point-inversion, positional symmetry, and separation. *Biological Cybernetics*, 51:305–312.

Freeman, W. and Adelson, E. (1991). The design and use of steerable filters. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906.

Fujita, I., Tanaka, K., Ito, M., and Cheeng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360:343–346.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269.

Gochin, P., Miller, E., Gross, C., and Gerstein, G. (1991). Functional interactions among neurons in inferior temporal cortex of the awake macaque. *Experimental Brain Research*, 84:505–516.

Gross, C. (1973). *Handbook of sensory physiology*, chapter Visual functions of the inferotemporal cortex, pages 451–482. Springer-Verlag, Berlin.

Gross, C. (1994). How inferior temporal cortex became a visual area. *Cerebral Cortex*, 4(5):455–469.

Gross, C., Bender, D., and Rocha-Miranda, C. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science*, 166:1303–1306.

Gross, C. and Mishkin, M. (1977). The neural basis of stimulus equivalence across retinal translation. In Harnad, S., Doty, R., Jaynes, J., Goldstein, L., and

Krauthamer, G., editors, *Lateralization in the Nervous System*, pages 109–122. Academic Press, New York.

Gross, C., Roche-Miranda, C., and Bender, D. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, 35:96–111.

Gross, C., Schiller, P., Wells, C., and Gerstein, G. (1967). Single-unit activity in temporal association cortex of the monkey. *Journal of Neurophysiology*, 30:833–843.

Hasselmo, M., Rolls, E., and Baylis, G. (1986). Object-centered encoding of faces by neurons in the cortex in the superior temporal sulcus of the monkey. *Soc Neurosci Abstr*, 12:1369.

Howard, J. and Kerst, S. (1978). Directional effects of size change on the comparison of visual shapes. *American Journal of Psychology*, 91(3):491–499.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148:574–591. London.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73(1):218–225.

Iwai, E. Mishkin, M. (1969). Further evidence on the locus of the visual area in the temporal lobe of monkeys. *Experimental Neurology*, 25:585–594.

Jolicoeur, P. (1987). A size-congruency effect in memory for visual shape. *Memory & Cognition*, 15(6):531–543.

Jolicoeur, P., Gluck, M., and Kosslyn, S. (1984). Pictures and names: Making the connection. *Cognitive Psychology*, 16:243–275.

Judge, S., Richmond, B., and Chu, F. (1980). Implantation of magnetic search coils for measurement of eye position: An improved method. *Vision Research*, 20:535–538.

Kimura, D. (1963). Right temporal-lobe damage. *Archives of Neurology*, 8:264–271.

Kubovy, M. and Podgorny, P. (1981). Does pattern matching require the normalization of size and orientation? *Perception & Psychophysics*, 30(1):24–28.

Lansdell, H. (1968). Effect of temporal lobe ablations on two lateralized deficits. *Physiol Behav*, 3:271–273.

Larsen, A. and Bundesen, C. (1978). Size scaling in visual pattern recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 4(1):1–20.

Legault, E. and Standing, L. (1992). Memory for size of drawings and of photographs. *Perceptual and Motor Skills*, 75:121.

Leung, T., Burl, M., and Perona, P. (1995). Finding faces in cluttered scenes using random labelled graph matching. In *Proceedings of the 5th Internatinal Conference on Computer Vision*, Cambridge, Ma.

Logothetis, N. and Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, 3:270–288.

Logothetis, N., Pauls, J., Bülthoff, H., and Poggio, T. (1994). View dependent object recognition by monkeys. *Current Biology*, 4(5):401–414.

Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563.

Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19:577–621.

Marr, D. (1982). *Vision*. Freeman.

Marr, D. and Nishihara, H. (1978). Representation and recognition of the spatial organization of three-simensional shapes. *Proceedings of the Royal Society of London. B*, 200:269–294.

McLain, D. (1974). Drawing contours from arbitrary data points. *The Computer Journal*, 17:318–324.

Milliken, B. and Jolicoeur, P. (1992). Size effects in visual recognition memory are determined by perceived size. *Memory and Cognition*, 20(1):83–95.

Milner, B. (1958). Psychological defects produced by temporal lobe excision. In *The Brain and Human Behavior*, volume XXXVI of *Proceedings of the Association for Research in Nervous and Mental Disease*. Williams and wilkins, Baltimore, MD.

Milner, B. (1968). Visual recognition and recall after right temporal-lobe excision in man. *Neuropsychologia*, 6:191–209.

Milner, B. (1980). Complementary functional specialization of the human cerebral hemispheres. In Levy-Montalcini, R., editor, *Nerve Cells, Transmitters and Behaviour*, pages 601–625. Pontificiae Academiae Scientiarium Scripta Varia, Vatican City.

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neuroscience*, 4:414–417.

Moody, J. and Darken, C. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294.

Nazir, T. A. and O'Regan, J. (1990). Some results on translation invariance in the human visual system. *Spatial Vision*, 5:81–100.

Pauls, J., Bricolo, E., and N.K., L. (1996). View invariant representation in monkey temporal cortex: Position, scale and rotational invariance. In Nayar, S. and Poggio, T., editors, *Early Visual Learning*, pages 9–41. Oxford University Press.

Perrett, D., Rolls, E., and Caan, W. (1979). Temporal lobe cells of the monkey with visual responses selective for faces. *Neurosci Lett Suppl*, S3:S358.

Perrett, D., Rolls, E., and Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, 47:329–342.

Perrett, D. I., Mistlin, A. J., and Chitty, A. J. (1987). Visual neurones responsive to faces. *Trends in Neuroscience*, 10(9):358–364.

Perrett, D. I., Smith, P. A. J., Potter, D. D., Mistlin, A. J., Milner, A. D., and Jeeves, M. A. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London B*, 223:293–317.

Pinker, S. (1984). Visual cognition: An introduction. *Cognition*, 18:1–63.

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

Poggio, T. and Girosi, F. (1990a). Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497.

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

Poggio, T. and Vetter, T. (1992). Recognition and structure from one 2d model view: observations on prototypes, object classes and symmetries. Technical Report A.I. Memo No.1347, Massachusetts Institute of Technnology, Cambridge, Ma.

Pouget, A. and Sejnowski, T. (1996). Spatial representations in the parietal cortex may use basis functions. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7, pages 157–164. MIT Press.

Rao, R. and Ballard, D. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence Journal*, 78:461–505.

Richmond, B., Optican, L., Podell, M., and Spitzer, H. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. i. response characteristics. *Journal of Neurophysiology*, 57(1):132–146.

Robinson, D. (1963). A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Trans Biomed Eng*, 101:131–145.

Rock, I. and DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology*, 19:280–293.

Rock, I., Wheeler, D., and Tudor, L. (1989). Can we imagine how objects look from other viewpoints? *Cognitive Psychology*, 21:185–210.

Rolls, E. and Baylis, G. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Experimental Brain Research*, 65:38–48.

Rosch, E., Mervis, C., Wayne, D. G., Johnason, D., and P., B.-B. (1976). Basic objects in natural categories. *Cognitive Psychology*, 6:382–439.

Rosenfeld, S. and Van Hoesen, G. (1979). Face recognition in the rhesus monkey. *Neuropsychologia*, 17:503–509.

Sanghera, M., Rolls, E., and Roper-Hall, A. (1979). Visual responses of neurons in the dorsolateral amygdala of the alert monkey. *Experimental Neurology*, 63:610–626.

Sáry, G., Vogels, R., and Orban, G. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science*, 260:995–997.

Schacter, D. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13(3):501–518.

Schwartz, E., Desimone, R., Albright, T., and Gross, C. (1983). Shape recognition and inferior temporal neurons. *Proceedings of the National Academy of Science USA*, 80:5776–5778.

Shepard, R. and Farrell, J. (1985). Representation of the orientations of shapes. *Acta Psychologica*, 59:104–121.

Shepard, R. N. and Metzler, J. (1971). Mental rotation of three dimensional objects. *Science*, pages 701–703.

Siegel, R. and Andersen, R. (1988). Perception of three-dimensional structure from motion in monkey and man. *Nature*, 331:259–261.

Sinha, P. and Poggio, T. (1994). View-based strategies for 3d object recognition. Technical Report A.I. Memo No.1518, Massachusetts Institute of Technnology, Cambridge, Ma.

Snodgrass, J. G. and Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117(1):34–50.

Tanaka, K., Saito, H.-A., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66:170–189.

Tarr, M. and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282.

Tarr, M. and Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science*, 1:253–256.

Tarr, M. and Pinker, S. (1991). Orientation-dependent mechanisms in shape recognition: further issues. *Psychological Science*, 2(3):207–209.

Taylor, L. (1969). Localization of cerebral lesions by psychological testing. *Clinical Neurosurgery*, 16:269–287.

Tenenbaum, J. and Bricolo, E. (1996). Analyzing the view dependence of population codes in inferior temporal cortex. To appear in CNS*96 proceedings.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32(3):193–254.

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

Ungerleider, L. and Mishkin, M. (1982). Two cortical visual systems. In Ingle, D., editor, *Analysis of Visual Behaviour*. MIT Press, Cambridge,Mass.

Vetter, T., Hurlbert, A., and Poggio, T. (1995). View-based models of 3d object recognition: Invariance to imaging transformations. *Cerebral Cortex*, 3(261–269).

Vetter, T., Poggio, T., and Bülthoff, H. (1994). The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology*, 4(1):18–23.

Von Bonin, G. and Bailey, P. (1947). *The Neocortex of Macaca Mulatta*. University of Illinois Press, Urbana, 4th edition.

Wachsmuth, E., Oram, M., and Perrett, D. (1994). Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. *Cerebral Cortex*, 4(5):509–522.

Wahba, G. (1990). *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia.

Weiss, Y. and Edleman, S. (1995). Representation of similarity as a goal of early visual processing. *Network: computation in neural systems*, 6(1):19–41.

Werkhoven, P. and Koenderink, J. (1993). Visual size invariance does not apply to geometric angle and speed of rotation. *Perception*, 22(2):177–189.

Yamane, S., Kaji, S., and Kawano, K. (1988). What facial features activate face neurons in the inferotemporal cortex of the monkey? *Exp Brain Res*, 73:209–214.

Yamane, S., Kaji, S., Kawano, K., and Hamada, T. (1987). Responses of single neurons in the inferotemopral cortex of the awake monkey performing human face discrimination task. *Neurosci Research*, S5:S114.

Young, M. and Yamane, S. (1992a). An analysis at the population level of the processing of faces in the inferotemporal cortex. In Squire, L., Ono, T., Fukuda, M.,

and Perrett, D., editors, *Brain Mechanisms of Perception and Memory: From Neuron to Behaviour*, pages 47–71. Oxford University Press, New York.

Young, M. and Yamane, S. (1992b). Sparse population coding of faces in the inferotemporal cortex. *Science*, 256:1327–1331.