Pergamon

0042-6989(94)00150-2

Vision Res. Vol. 35, No. 4, pp. 549–568, 1995
Elsevier Science Ltd. Printed in Great Britain

Object Classification for Human and Ideal Observers

ZILI LIU,* DAVID C. KNILL,† DANIEL KERSTEN‡

Received 15 October 1993; in revised form 4 July 1994

We describe a novel approach, based on ideal observer analysis, for measuring the ability of human observers to use image information for 3D object perception. We compute the statistical efficiency of subjects relative to an ideal observer for a 3D object classification task. After training to 11 different views of a randomly shaped thick wire object, subjects were asked which of a pair of noisy views of the object best matched the learned object. Efficiency relative to the actual information in the stimuli can be as high as 20%. Increases in object regularity (e.g. symmetry) lead to increases in the efficiency with which novel views of an object could be classified. Furthermore, such increases in regularity also lead to decreases in the effect of viewpoint on classification efficiency. Human statistical efficiencies relative to a 2D ideal observer exceeded 100%, thereby excluding all models which are sub-optimal relative to the 2D ideal.

Object recognition Ideal observers Template matching Radial basis functions

1. INTRODUCTION

Accurate object recognition and classification is crucial for humans and animals to successfully interact with their environment. These functions include a number of aspects of object perception: The determination of what category an object fits into, how well it fits into a selected category, and the discrimination of members of the same category from one another. Though much research, both computational and experimental, has been done in the area, the nature of the brain processes and representations subserving object recognition remain an area of strong debate.

An important first step in understanding human object recognition is to understand the functional aspects of human performance, i.e. to develop a computational theory of how the human visual system performs the various tasks associated with object recognition. Such a theory will provide fundamental constraints on models of the architecture and processing characteristics of the recognition system(s) built into human vision. As suggested by Tarr (1992), computational questions about object recognition can be classified into a number of roughly independent categories: (1) the spatial dimensions (2D or 3D) which characterize object representations; (2) the frame of reference or coordinate system in which an object is specified, e.g. viewpoint dependent vs viewpoint independent; (3) the nature of object component parts; and (4) the spatial relations between these parts. To this list we add a fifth: (5) the regularities in object geometry (e.g. symmetry) which are taken advantage of in both object representations and the processes which match image data to these object representations for recognition.

In this paper, we present a novel experimental paradigm, based on ideal observer theory, to address some of the issues posed above. The development and application of the ideal observer paradigm to a problem in high-level perception such as object recognition is a major part of the focus of the paper. The approach provides a formal means for making strong inferences based on the quantitative performance of subjects in an experimental task. The particular empirical questions we will investigate are: do internal models of objects and the matching processes subserving recognition incorporate only 2D, image-based information or do they include 3D descriptions? Do object representations and the matching process make use of special regularities in the shapes of objects such as symmetry? Is object classification performance better for previously seen views of an object than for novel views (i.e. is it viewpoint dependent)?

1.1. Metric Object Recognition

When considering object recognition, a particularly useful distinction is between the categorization of objects based on qualitative differences in shape (the difference between a car and a chair), and the categorization of

*NEC Research Institute, Princeton, N.J., U.S.A.

†Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

‡To whom all correspondence should be addressed at: Department of Psychology, 75 East River Road, University of Minnesota, Minneapolis, MN 55455, U.S.A. [*Email* kersten@eye.psych.umn.edu].

objects based on metric differences in shape (the difference between a football and a basketball).* The distinction is useful for two reasons. One is that an object's functional utility to an organism is often determined by its shape so that the distinction between qualitative and metric differences in shape often maps nicely onto a distinction between qualitative and quantitative differences in functional properties of objects. A second is that qualitative differences in object shape are often reflected in invariant qualitative differences in their images (e.g. topology of object boundary contours), while metric differences are generally reflected in viewpoint-dependent metric differences in the images (Biederman, 1987). This opens the possibility that quite different strategies might be used for the different levels of object categorization and discrimination (but see Edelman, 1991).

We use an experimental task in which subjects make classification judgments based on metric differences in object shapes; therefore the results reflect those aspects of the human visual system which subserve metric object recognition. In this paper, we will use the term object recognition to refer to the classification and discrimination of objects based on their metric shape properties.

1.2. Computational Theory of 3D Object Recognition and Discrimination

A system for recognizing, classifying, or otherwise comparing image information about objects with internal models of objects in memory must have three components, as diagrammed in Fig. 1. These include a process which produces some representation of the input data, an internal model, and a process by which the input and internal models are matched. Multiple stages of such systems are possible [e.g. a hierarchy of qualitative and quantitative matching processes (Biederman, 1987)], and the discussion to follow generalizes to such multiple stage processes. Most theories of human object recognition do not consider the actual process used for coding the input representation (involving, for example, segmentation and feature extraction), but start with a specification of the representation itself. We will, by and large, follow this tradition here and we take a basic recognition system to contain three components, the input representation, the internal model and the matching process. In all that follows, we will avoid the problem of self-occlusion and assume that none of the views obscure any of the feature points. In principle, one can deal with the problem of self-occluding parts by restricting the analysis to each node of an object's aspect graph (Koenderink & van Doorn, 1976), i.e. one applies the analysis to sets of views that do not lead to qualitative changes in the object's image. In practice, self-occlusion is a non-trivial issue because the computation of an object's aspect graph can be difficult (Plantinga & Dyer, 1990).

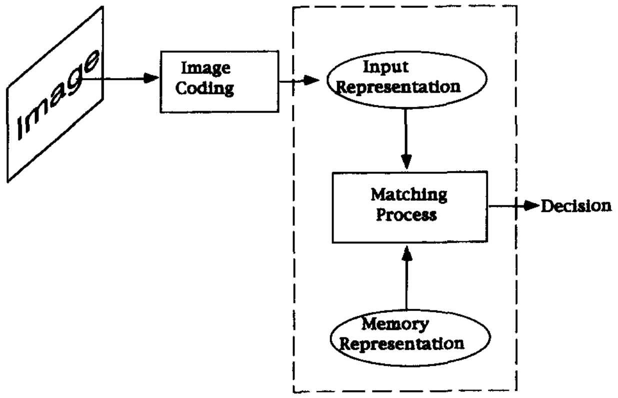

FIGURE 1. Schematic diagram of a general processing system for comparing image data with internal object representations for purposes of recognition. The components of the system outlined in the dashed box are those considered as comprising an object recognition system.

Object recognition theories differ in the characteristics assigned to each of the three components of the recognition process. The broadest dimensions along which the models differ are in whether the input and internal object models are assumed to be 2D representations of the images of objects (e.g. 2D positions of object features) or 3D representations of the objects' shapes. Strongly related to this is a similar distinction made about the matching process. Does the matching process incorporate 3D knowledge about objects which is not explicitly included in the internal representation? Importantly, from the point of view of absolute system performance, the above two distinctions may be confounded. Consider the following two architectures. In one, internal object models are 3D representations, the input is a 2D description of one view of an object, and the matching process checks the input against all possible views of its stored object models in order to determine which object is being viewed. In the other, the internal representations are sets of 2D descriptions of different views of objects, the input is again a 2D description of a view of an object, and the matching process has knowledge of the constrained nature of views of rigid objects;† namely, that all views fall on a well-defined hyper-surface in the space of 2D representations which is fully determined by just two views (Poggio & Vetter, 1992; Ullman & Basri, 1989).

*The distinction between qualitative and metric differences in object shape is not entirely clear cut. In order to objectively make such a distinction, we would have to fall back on mathematics, e.g. by equating qualitative with topological. Ultimately, for human object recognition, it is a psychological question, and relies on an assumption that the perceptual system represents some object attributes categorically and others along approximately continuous, ordered scales. The qualitative/metric distinction is similar to the basic level/subordinate level distinction often made for different types of linguistic categories (Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976). While the latter distinction is operationally (and somewhat fuzzily) defined by the psychological characteristic that basic level categories are ones for whom names readily come to mind, the two types of categories may often also be distinguished by the fact that categories at the basic level differ in the types of features they have, while those at the subordinate level differ in the values assigned to the features.

†With enough stored views, some non-rigid transformations (e.g. non-rigid affine transformations) of objects can also be allowed in the matching process.

Since this constraint means that the views which the system has stored for each object fully determine the legal views of the object, just as a 3D object model would, a system storing object models as 2D view descriptions in conjunction with an appropriate matching process could perform as well as one with 3D object models. We would say that such a system would have 3D information incorporated into the matching process. Keeping this consideration in mind, we distinguish between four types of object recognition systems, each of which would give qualitatively different performance: (1) 2D/2D systems, in which both internal object models and input data are 2D descriptions of specific views of objects, and in which the matching system does not know the constraints relating one view of a rigid object to another, but simply matches novel views to stored views; (2) 3D/2D systems, in which the input is a 2D description of an image and either the internal object models are 3D descriptions of an object's shape or 3D constraints are built into the matching process; (3) 2D/3D systems, in which the internal models are 2D descriptions of learned views of objects, and the input is a 3D description of object shape, inferred from the image data; and (4) 3D/3D systems, in which both input and internal representations are 3D descriptions of object shape.

A second distinction which one can make about both the representations and the matching process used in an object recognition system is whether they are viewpoint dependent or not (Bülthoff & Edelman, 1992; Tarr & Pinker, 1989). Naturally, all 2D representations are themselves viewpoint dependent, but, as described in the example above, having 2D internal models by itself does not necessarily lead to viewpoint dependent performance, if the matching process does a sufficiently good job of determining from the stored views the constraint surface relating all possible views of an object. Viewpoint dependent performance arises from how the matching process compares input data to stored object representations. The 2D/2D system described above is inherently viewpoint dependent, while the others can vary from being strongly viewpoint dependent to being completely viewpoint independent. In either a 2D/3D or 3D/2D system, viewpoint dependence can arise from having a viewer-centered object model and a matching process whose ability to align the model with the input data is limited. This could arise from limits on the possible angles of internal rotation or on uncertainty which is effectively added by the internal alignment process, increasing as the object model is taken through successively larger angles in the alignment. This suggests that while a finding of viewpoint independence would effectively eliminate one class of models, a finding of viewpoint dependent performance does not by itself distinguish among the different classes of recognition system described above.

A final issue with which we will be concerned is that of what types of object regularity the human recognition system makes use of in representation or matching. Rock and DiVita (1987) have suggested that very irregular objects may be represented using 2D image descriptions and then matched in a straightforward way with novel input images, whereas regular (e.g. symmetric) objects may be represented using a 3D description of shape. They suggest that such a dichotomous representation/matching scheme could result simply from efficiency considerations; namely, that while a 3D representation is significantly more efficient for representation of a regular object, no such benefits accrue for irregular objects, which may be almost as efficiently represented as a set of 2D views. Special mechanisms for the encoding of regular 3D shapes could also exist as a result of the preponderance of symmetric objects in our world and the arguably greater functional significance of such objects to humans. Another possibility is that special mechanisms exist for matching models of regular objects to novel views. Poggio and Vetter have shown that matching symmetric objects requires less information about an object (fewer stored views) than does matching irregular objects and have demonstrated the computational (Poggio & Vetter, 1992) and psychological (Vetter, Poggio & Bülthoff, 1994) feasibility of mechanisms which incorporate the appropriate constraints. The potential payoffs of having an object recognition system designed to take advantage of object regularities suggest that the human visual system may well incorporate such design strategies.

A number of experiments using reaction time measures have shown differences in the viewpoint dependency of object recognition performance between different types of objects. The time to perform an object recognition task for objects constructed from several different qualitatively different parts ("geons") is not significantly affected by object rotation in depth (Biederman & Gerhardstein, 1993), whereas reaction time is affected by such rotations for objects built from qualitatively similar parts (Gerhardstein, 1992). At this point, however, we have few systematic studies of the role of object regularity in recognition. It is not clear whether there are general forms of object regularity, such as symmetry, which reduce the need for multiple views in recognition or whether it is the presence of parts with non-accidental contrasts. Although we do not address this specific question here, we will investigate the issue of whether object regularity changes how well subjects recognize objects and whether it changes the viewpoint dependency of overall performance on a recognition task.

1.3. Models of Object Recognition

A large assortment of object recognition models have been developed over the last 30 yr. We will not attempt to review all of them here, but rather focus on a few recent proposals that are most directly applicable to the problem of metric object recognition. Models such as Biederman's *recognition-by-components* model (Biederman, 1987) are directed more toward qualitative recognition. We describe two types of model, those which incorporate alignment mechanisms and those based on associative learning of images and object representations. The two types of model are not mutually exclusive, but the distinction forms a working organization for the discussion. We will not present critiques of the different types of mode, but rather focus on categorizing the models within the representation-based taxonomy presented above.

Alignment schemes, which have been generally characterized by Huttenlocher and Ullman (Huttenlocher & Ullman, 1987; Ullman, 1989), can fit into any of the four classes of models, depending on the nature of the representations assumed. In alignment models, either the input is first aligned with the internal object models, i.e. is transformed (rotated, scaled, etc.) until it best matches each of the models; or the object models are aligned until they best match the input. After alignment, the input is matched to the object models, and the model which it best matches is chosen for the object categorization. That one or the other of the input and internal representations is viewpoint dependent is explicitly assumed in alignment models (hence the need for an alignment process). In fact, most alignment type models use viewpoint dependent 3D object models and 2D image inputs, thus are of the 3D/2D type (Huttenlocher & Ullman, 1987).

Associative neural network models have been applied to a range of problems in 2D pattern recognition. These models have typically been developed to deal with invariant recognition over 2D transformations and would thus fall within the 2D/2D class of systems (Anderson, Silverstein, Ritz & Jones, 1977; Bienenstock & Von der Malsburg, 1987; Kohonen, 1978). Only recently have there been associative learning models that attempt to handle rotation in depth (Poggio & Edelman, 1990). In Poggio and Edelman's *generalized radial basis function* (GRBF) model, multiple views of an object are associated with a canonical view of the object. When a new view of an object is presented to the system, the system transforms it into a representation which is then matched to the learned canonical view and is accepted as being a view of the object based on the degree of fit between the reconstructed view and the canonical view. The learned transformation has built into it both the object models learned (the training views) and the process for matching them to new views. The GRBF model, as it has been applied to object recognition, uses 2D input representations and appears to use 2D internal models, as it transforms the input to a new view of an object. However, we note that the nature of the output representation is of little importance to the model (see Appendix D), and the internal model should most naturally be considered to be the learned parameters of the network. Since the system essentially learns to approximate the mapping from a wide range of views to the same output, it effectively is approximating the constraint surface relating views of an object to one another. The distinction between representation and process is lost in an associative system like the GRBF model, so that we cannot neatly delineate whether 3D information is contained in the representation or the matching process, however it is clear that 3D information is contained in the system, leading us to classify it as a 3D/2D system.

2. THE IDEAL OBSERVER APPROACH: MEASURING EFFICIENCIES FOR HUMAN OBJECT RECOGNITION

We are particularly interested in designing experiments geared toward answering questions about the computational aspects of high-level human visual processing, such as what information does the visual system use to perform a task, what internal representations does it use and what general processing strategies does it use? A number of approaches are typically applied to the psychophysical investigation of these questions, but we can broadly categorize them into two classes: those looking at how long it takes subjects to perform a task (reaction time studies) and those looking at how well subjects perform a task (e.g. measuring percentage correct, discrimination thresholds, etc.). The former converts hypotheses about system architecture, representation schemes, etc. into hypotheses about the time–course of performance for a task by making assumptions about the temporal nature of visual system processing. The latter class of experimental paradigms tests hypotheses about overall system performance. In this discussion, we will focus on the latter approach.

Among the paradigms based on measuring overall performance, we can distinguish between two types. In one approach, theories about the computational aspect of a perceptual function such as object recognition are translated into hypotheses about the qualitative behavior of subjects on a psychophysical task. In particular, predictions are made about the relative performance of subjects for different classes of stimuli or for different tasks (e.g. learned vs novel views). A common problem with this approach is the lack of a "common currency" for comparing performance across different stimuli or different tasks. In order to assess relative performance, one needs a measure of performance which accurately reflects subjects' internal processing characteristics and constraints. It should not confound effects due to such internal processing characteristics with those due to differences in the information available in different classes of stimuli for the performance of a task or in the information available for the performance of different tasks. It is quite possible that the causes of differences in performance across different stimuli or tasks (e.g. regular vs irregular objects) are in the stimulus, not in the head.

A second approach to testing theories using overall performance on a psychophysical task is to compare the performance of particular models to that of human subjects. In a recent example of this, Bülthoff and Edelman (1992) compared performance curves on an object recognition task as a function of orientation of novel views away from learned views with a similar performance curve for the GRBF model of recognition. They found that the pattern of decreased performance with increasing angular difference between learned viewpoints and test viewpoints was very similar for their subjects and the model. Such results, however, do not allow us to test what aspects of the computational theory on which the model is based accurately reflect human visual processing. Most models, like GRBF, are constrained both by computational theory and by implementation constraints, thus their performance is a function of both the computational theories on which they are based and the particulars of the implementation (e.g. using Gaussian interpolation functions in the case of GRBF). In the example just described, one is left with a new problem. What properties of the model give it viewpoint dependent performance, and are these the same things which give human subjects viewpoint dependent performance?

The ideal observer paradigm speaks to both of the problems described above. It allows one to make quantitative predictions of performance based on theories expressed at the computational level of description and provides a common measure of performance with which to compare subjects' performance across different types of stimuli and different tasks. The basic idea behind the ideal observer paradigm is that for any psychophysical task, an ideal observer can be defined which performs the task as well as the stimulus information allows, i.e. the ideal observer provides a theoretical ceiling on performance, which other observers can approach but cannot exceed (Kersten, 1990). Moreover, different ideal observers can be built to match different computational constraints on a system, e.g. ideal observers can be defined for an object recognition task in which one ideal has access to a full 3D model of learned objects and the other only has access to a set of learned 2D images of an object. These ideals are optimal implementations of different computational theories in the sense of giving the best possible absolute performance of any implementation of the theories (they may, however, be much too slow for practical application). Naturally, if subjects perform better than a constrained ideal, then one can infer that the computational constraints on that ideal are not built in, in their entirety, to the human visual system.

The ideal observer is formulated in terms of Bayesian statistical inference. For example, in an object recognition task, an ideal observer would select, from a set of candidate object models, the one which is most likely to give rise to the input image. In Bayesian terms, it would select an object model which maximizes the posterior probability:

$$P(\text{object}\,|\,\text{image}) = \frac{P(\text{object})P(\text{image}\,|\,\text{object})}{P(\text{image})} \quad (1)$$

where $P(\text{object}\,|\,\text{image})$ is the conditional probability of the object given the input image; $P(\text{object})$ is the prior probability of the object; $P(\text{image}\,|\,\text{object})$ is the conditional probability, or likelihood, of the input image given the object; and $P(\text{image})$ is the probability of the input image. Since the comparison is made across objects and $P(\text{image})$ is constant for a fixed input image, we do not need to know the distribution of images. When the prior probability, $P(\text{object})$ is constant, maximizing the posterior probability is equivalent to maximum likelihood estimation.

Ideal observers allow one to define a measure of the statistical efficiency with which human subjects use the information available for performance of psychophysical tasks (Barlow, 1980). Efficiency gives a measure of performance which is relative to the amount of stimulus information available for a task, and as such provides a common dimension for comparison across different stimuli and different tasks. Typically, efficiency is measured in terms of the relative signal-to-noise ratios needed by ideal and human observers to achieve the same level of performance (noise, as the term is used here, refers to stimulus uncertainty, whether it is inherent in the task or is artificially induced by the addition of stimulus noise in an experiment). It is immediately clear that tasks for which the ideal is perfect do not support the computation of useful efficiency measures, since for such tasks efficiency is guaranteed to be 0, or at best, if humans are also perfect, undefined. An important aspect of experimental design within the ideal observer paradigm, therefore, is that the visual task used have some inherent uncertainty. This uncertainty can arise naturally due to loss of 3D information in the image or can be added artificially in the form of noise added to the stimuli. Both kinds of uncertainty are incorporated in the experimental task described below.

We use ideal observers in two ways in this paper. First, we construct an ideal observer for the object classification task we use based on the assumption that the observer uses only 2D information to perform the task. The ideal bases its judgments on a comparison between a stimulus image and a set of 2D learned views of an object; thus, it is the ideal for the 2D/2D class of models described in the previous section. If subjects are able to attain greater than 100% efficiency relative to this 2D/2D ideal, then we can eliminate that class of recognition models as possible models for the task performed, i.e. we can infer that subjects incorporate some 3D knowledge of either the object in the stimulus or the object stored in memory, or both. Second, we construct an approximation to the "true" ideal for the task, which provides a measure of the best possible performance on the task. This ideal serves as an absolute benchmark for performance in different conditions of the experiment. The ideal we use for this purpose is what we refer to as the 3D/2D ideal. It builds up a complete 3D model of a learned object from the training views and matches it against the 2D images presented in the experiment.* Finally,

*Some very coarse 3D information is available in the stimuli we use in the form of occasional occlusions. Such occlusions provide relative depth information. Though incorporation of this information would slightly improve the performance of the ideal, we think these improvements would be so small as to have no significant effect on the interpretation of results.

we derive a 3D/3D ideal observer which not only has an accurate 3D internal model of objects, but also is able to accurately infer the 3D structure of objects in stimulus images. This ideal is a super-ideal, in that the 3D information it uses as input data is in large part not available to human observers in our experiment. It does, however, serve as a benchmark against which to compare the other ideals to see how different computational constraints affect performance.

We have noted that absolute differences in subjects' performance in different experimental conditions reflect differences both in the information content and in the internal processing of the stimuli. The differences in efficiency relative to the 3D/2D ideal, on the other hand, reflect only differences in internal processing of the stimuli. A particularly important consideration in this regard is that performance in any recognition task depends on the choice of distractors. Effects due to differences in distractor sets under different conditions of an experiment are effectively absorbed in an efficiency measure of performance. Suppose, for example, that in one condition the distractor set of objects is more confusable with the object to be recognized than in another condition. This might, by itself, lead to worsened subject performance in the first condition, however the effect would also show up in the performance of the ideal observer. If the change in subjects' performance was entirely due to the difference in confusability of the distractor set, the relative efficiency of subjects' performance in the two conditions would remain unchanged.

To compare the performance between human and ideal observers experimentally, we have chosen 3D wire objects (Fig. 2) for our classification task. The wires are created by joining a set of 3D vertex positions with thick cylindrical wires. In the experiment, subjects are asked to decide which of the two images of noisy versions of a learned prototype wire object is more similar to the prototype. By noisy wires, we mean wires whose vertices have been disturbed by uncorrelated positional noise. Four reasons for the use of wire objects in this study are: (1) experiments using the same kind of wire objects have been used (Bülthoff & Edelman, 1992; Poggio & Edelman, 1990) to support a 2D view interpolation theory of object recognition; (2) the novel wire objects are suitable for addressing the nature of internal representations and viewpoint dependency, since we need not worry about subjects' prior familiarity with the objects (Rock & DiVita, 1987); (3) the corresponding ideal observer models are computable, since occlusions of the feature points (vertices) are negligible; and (4) thick wires can be rendered to appear as 3D objects, with the partial occlusion at the vertices indicating relative depth. Although the use of thick wire objects might limit the generality of our results, these objects are nevertheless useful for studying the internal representations of spatial relations between object features.

3. METHODS

3.1. Stimuli

The stimulus images were derived from 12 different prototype wire objects. The prototypes were divided into four types, each consisting of three objects (Fig. 2) and which varied in the degree and type of regularity. The four types were: Balls—five small balls randomly positioned in 3D; Irregular wires—the three sets of five randomly placed balls such as used as objects in the Balls category were connected by four straight cylinders of the same diameter as the balls, making three thick wire objects with the appearance of bent paper-clips; Symmetric wires—wire objects were made as above with the constraint that the objects were mirror symmetric around an imaginary plane slicing the middle of the objects; V-Shaped wires—similar to the Symmetric wires with the added constraint that the two cylinders on each side of the plane of symmetry were colinear, so that the objects formed a V-shape, making them both mirror symmetric and planar. Each object was constructed by specifying the 3D positions of the five balls. The distance between neighboring balls in a prototype object was always 3.55 cm (approx. 80 pixels when in the image plane). The positions of the succeeding balls in an object were specified in spherical coordinates relative to the position of the previous ball. The angle with the z-axis was chosen from a uniform distribution between 0 and 180 deg and the angle with x-axis was chosen from a uniform distribution between 0 and 360 deg. The geometric center of the object was then calculated from the ball positions in 3D and moved to the origin. Each ball had a diameter of 0.23 cm.

Stimulus images were created by orthographic projection of these objects onto the screen plane of the computer display. The Doré 3D graphics package on a Stellar ST2000 computer was used to render the wire objects. The objects were rendered with a matte Lambertian reflectance. The light source was modeled as a point source at infinity, with tilt 0, and slant 63.44 deg. Shading indicated to the observers that the images were of 3D configurations, rather than 2D patterns. For the wire objects, the shading pattern provided some clue as to the relative slant of the wire segments. It also caused contrast edges (T-junctions) to be formed at self-occlusions, providing further ordinal depth information. Size changes were absent, because of orthographic projection.

In each experimental session, subjects were trained on 11 different views of one of the prototype objects. The 11 training views of a prototype object were created by rotating the object first around the x-axis (horizontal in the screen plane) six times in 60 deg steps, and then around the y-axis (vertical in the screen plane) six times, again with 60 deg rotational steps (see Fig. 3), resulting in 11 views of the object. The center of the object (defined as the geometric center of the five vertex positions) was used as the origin for the rotational movement of the object. The training views of an object extended over an average range of approx. 5 cm on the

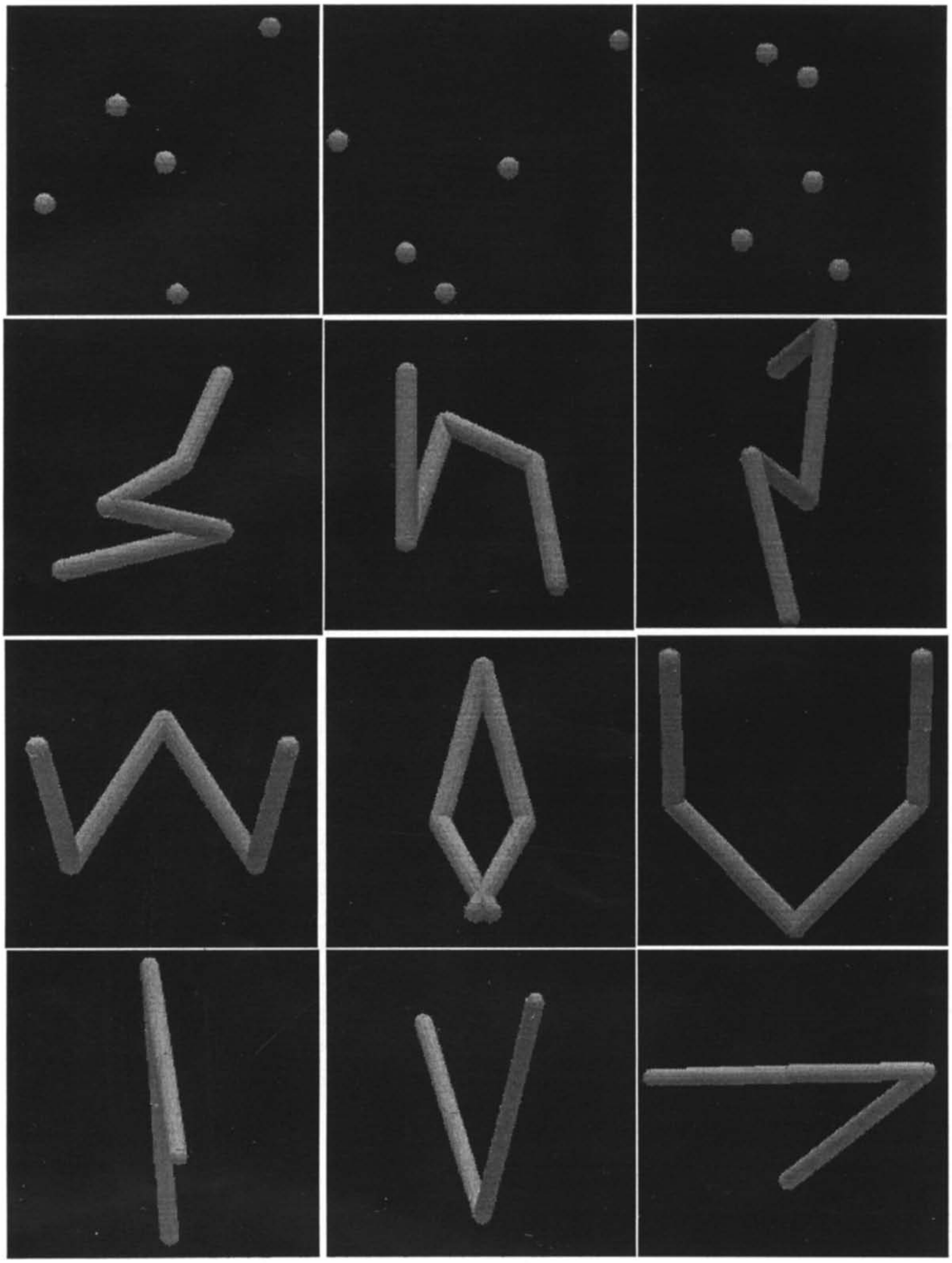

FIGURE 2. Objects used in the psychophysical experiments. Top row, Balls; second row, Irregular; third row, mirror Symmetric; bottom row, V-Shaped.

screen. With the viewing distance of 150 cm used, this corresponded to approx. 1.9 deg of visual angle.

On each trial of the testing phase of an experimental session, two new noisy versions of the prototype objects were created and projected onto the computer screen. A fixed amount of uncorrelated Gaussian noise (SD = 0.254 cm) was added to the vertex positions of the learned prototype to create a *target* object. A variable amount of noise, with a greater SD than that of the *target* object, was added to the vertex positions of the prototype to create a *distractor* object. The two test objects were then orthographically projected to the computer screen from the same viewpoint. On half of the trials, the viewpoint was selected randomly from a

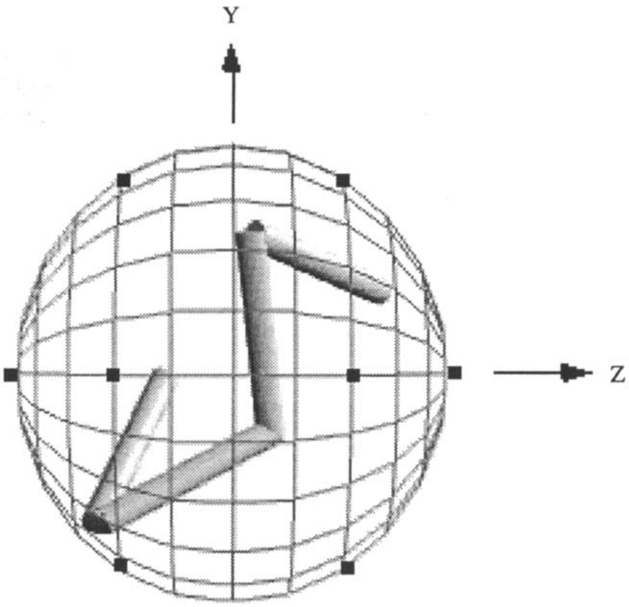

FIGURE 3. The 11 template views are indicated in this viewing sphere by the black squares. Subjects view the object along the z-axis. There are two more views on the equator behind, which are occluded. The rightmost square represents two views which differ by 180 deg of rotation around the z-axis.

uniform distribution over the viewing sphere. On the other half of the trials, one of the 11 training views, randomly chosen, was used. This created two conditions for testing of each prototype object, an old-view condition and a new-view condition. Because of the noise added to the prototype object, the variance in the visual angle subtended by the stimuli was greater than for the learned views.

3.2. Apparatus

Stimuli were presented on the CRT of the computer's console. Subjects made responses by pressing one of two buttons on the computer's mouse. We wished to simulate, as close as possible, the viewing of the real 3D objects. Subjects viewed the stimuli monocularly through a hole cut in the face of one end of a wooden tube whose inside was painted black. The other end was open and abutted the CRT. The background screen color for the stimuli was black, and the border of the computer screen was covered with black cardboard. The chin rest from which subjects viewed stimuli was 13 cm from the hole, so there was virtually only one position from which subjects could see the stimuli, effectively eliminating motion parallax as a flatness cue.

3.3. Procedure

Subjects ran a total of 12 sessions, one for each prototype object, grouped into three blocks. The four sessions in any one block tested prototype objects from each of the four types used in the experiment. The order of testing the prototypes was randomized across subjects. Each experimental session consisted of three phases:

3.3.1. Learning

Subjects were shown an ordered list of the training views as described above (Section 3.1).

3.3.2. Practice

Subjects were presented with two stimulus images shown side by side on the CRT. One of the stimuli was a training view of the prototype (chosen randomly from the list of 11), the other was a view of a distractor object. The distractor object was generated either by adding noise to the vertex positions of the prototype object or by random generation of a new wire object. Equal numbers of the two types of distractor were used. Target and distractor stimuli were presented equi-distant from the center of the screen along the horizontal axis. The left/right positions of target and distractor stimuli were random between trials. Subjects indicated which of the two stimuli was one of the training views by pressing the left or right buttons on the mouse. Feedback was given in the form of erasing the image of the distractor (leaving the target on the screen) after each trial. If a subject's response was incorrect a bell sounded. The criterion for moving from the practice phase of the session to the test phase was that the observer responded correctly on 45 out of the most recent 50 trials. The maximum number of practice trials allowed was 100.* At any time, subjects could press a third mouse button to review the training stimuli, which would be presented in order to the subject. Subjects controlled the rate of review using the mouse.

3.3.3. Test

During the test phase, target and distractor stimuli were shown side by side on the screen and subjects were instructed to choose the one which was most similar to the learned prototype by pressing the left or right mouse button. Both stimuli in a given trial were generated from the same viewpoint. Figure 4 illustrates the stimuli for the task, and more details are given in Section 3.4. For 40% of the trials, the stimuli were generated from the same viewpoints used to generate the training stimuli. We refer to this as the Old-view condition. For another 40% of the trials, the stimuli were generated from novel viewing positions (randomly drawn from a uniform distribution over the viewing sphere). We refer to this as the New-view condition. The remaining 20% of trials were repeats of the trials used in the practice phase of the session. These were used to maintain a subjects' memory of the learned prototype. In these trials, the target stimulus was a training view of the prototype (with no noise added) and the distractor was the same view of a noisy version of the prototype. We refer to these trials as *maintenance* trials. No feedback was given for the test trials. Feedback was given for

*Subject DNB did not reach a 90% hit rate at the end of the practising phase for the first two Balls objects and the first Irregular object, neither did subject GLJ for the first Ball object. But their hit rate performance was within the range 80–89% for these objects during their last 50 trials.

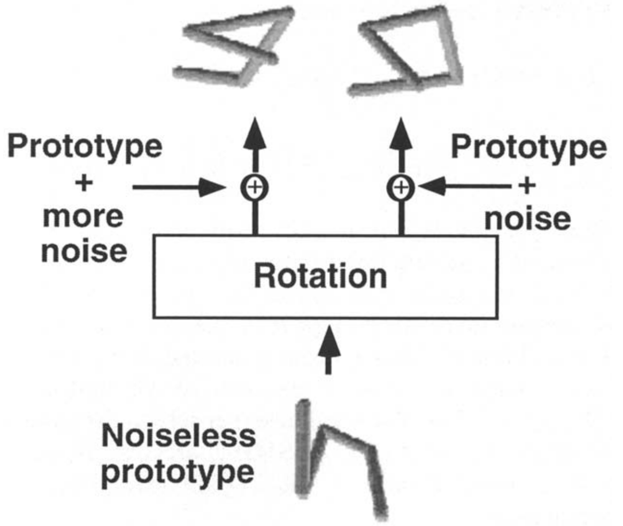

FIGURE 4. In the test phase of the experiment, two stimuli were generated by a 3D rotation of the noiseless prototype. The standard deviation of the positional noise added to the distractor (prototype + more noise) was greater than that added to the signal (prototype + noise).

maintenance trials. Subjects could review the prototype object (i.e. its 11 template images) at any time, as described above for the practice phase of the session. All trials were randomly ordered. The test phase consisted of 300 trials, about 240 of which were test trials (about 120 old views and about 120 new views) and about 60 of which were maintenance trials.* Subjects were allowed to take an optional break after every 50 trials. We used the QUEST (Watson & Pelli, 1983) adaptive staircase procedure to find the distractor noise level needed for subjects to perform at a 75% correct response level. Thresholds were estimated independently for the Old-view and New-view conditions. Three paid, naive subjects with normal vision participated in the experiment.

3.4. Ideal Observers and Models

3.4.1. The task

For the ideal observers, wire objects were represented as ordered sequences of vertex coordinates. This effectively assumes known correspondence during recognition. In the ideal observer models, we also allowed for an ambiguity regarding which of the two ends was the beginning. The 2D/2D ideal represented both the prototype object to be matched and the stimuli presented on a trial as ordered sets of 2D coordinates of object vertices in an image. The prototype object was represented as 11 sets of vertices corresponding to the 11 different views of the prototype learned by subjects in the experiment. The 3D/3D ideal represented both prototype and stimulus objects by their 3D vertex coordinates. Note that for the 3D/3D ideal, we have assumed that the 3D positions of the vertices of each noisy object could be accurately extracted from the image by the ideal. In this sense, the 3D/3D ideal is a super-ideal, since it does not take into account any error in the perceived 3D shape of objects in the stimulus images.

The task for the ideal was the same as for the human subjects—given a prototype wire object and images of two noisy versions of the prototype, decide which of the two noisy stimuli was generated with the least noise. The logic of the task is illustrated best using the 3D/3D ideal. The prototype can be represented as a point in a 15 dimensional space (5 vertices × 3 coordinates). The target stimulus is generated by adding uncorrelated Gaussian noise with fixed variance to the vertex positions of the prototype and the distractor is generated by adding uncorrelated Gaussian noise with more variance to the vertex positions. The ideal has to decide which of the two stimuli is the target. In effect, it has to select which of the two stimuli is most similar to the prototype in the sense of minimizing the mean squared distance between the vertex positions of the stimuli and the prototype. This signal discrimination task for one dimension is illustrated in Fig. 5. Hence the ideal will make a mistake when a sample point from the distractor distribution is closer to the prototype than a sample point from the target distribution.

Thresholds for the ideals were measured as the increase in noise one needed to add to the distractor over that added to the target to make the classification decision correctly 75% of the time.

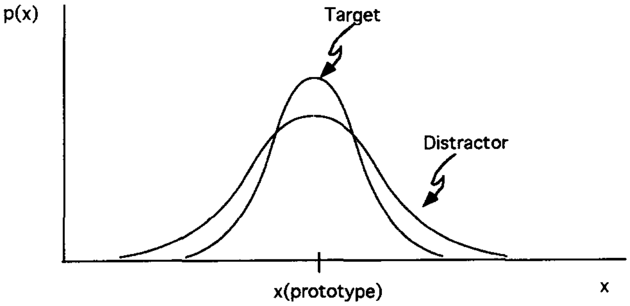

FIGURE 5. The signal discrimination task presented to ideal observer. The ideal must discriminate between two classes of wire object—one generated by adding a small amount of noise to a prototype object (the target), and the other generated by adding a larger amount of noise to the prototype object (distractor). The means for both target and distractor sets are the same prototype object.

*The experimental design changed after running subject GLJ. For this subject, an equal number of test and maintenance trials was used, resulting in a total of 480 trials instead of 300, and an optional rest at every 80 trials instead of 50. For this subject, test and maintenance trials were presented in regular alternation. For subject GLJ, feedback was predictable. The subject reported that he paid more attention to the trials with feedback, unaware of the experimental purpose. We believe that since old and new views were randomly mixed and only trials without feedback (test trials) were counted, the subject still used the same strategy in the test for new and old views. Moreover, the subject's absolute level of performance on test trials was not particularly different from those of the other two and the earlier pilot subjects. We, therefore, include GLJ's data with the rest of the results.

3.4.2. *The ideals*

We simulated four different ideals and one model for comparison with human subjects' performance. Three of the ideals and the model (GRBF) were introduced in Section 2. The three ideals are the 2D/2D ideal, the 3D/2D ideal and the 3D/3D ideal. To these three we add a fourth based on the observation that learning of an object could take place during the course of the test trials, not just the training trials. Such learning would not affect the 3D/2D or 3D/3D ideals, since these already have assumed an accurate and complete 3D representation of a prototype object has been learned in training. The 2D/2D ideal, however, could potentially be improved by including test stimuli as templates to be matched. Though these are noisy versions of the prototype, learning them would approximate learning new views of the prototype. We call this ideal the Learning 2D/2D ideal.

3.4.2.1. 2D/2D ideal observers. The vertex positions of the target stimulus image in each trial are generated by first adding Gaussian noise independently to each of a prototype's vertex coordinates in 3D and then projecting those vertices orthographically onto the image plane. Orthographic projection amounts to simply removing the z-coordinate of each vertex, so that it is equivalent to a process which first projects the prototype vertices onto the image plane and the adds independent Gaussian noise to the 2D vertex positions of the resulting wire image. The 2D/2D ideal observer matches stimuli against all 2D rotations of the 11 learned template views of the prototype. This can be seen as a sampling of the space of views available to the 3D/2D ideal. Accordingly, the 2D/2D ideal effectively assumes that the target stimulus was generated by adding a fixed level of noise to any one of the 11 learned templates at any of the possible 2D rotations of the image. In the limit, as the number of learned views approaches infinity (covering the viewing sphere), the performance of the 2D/2D ideal will approach that of the 3D/2D ideal. For the human and ideal observers, noise was not added to one terminal vertex of the object. So for the ideal, during each trial, a terminal vertex of a template was always aligned with a terminal vertex of a stimulus. We assumed that the 2D/2D ideal knew the sequence of input vertex coordinates, but did not know if a terminal vertex was the beginning or end of the sequence. This last assumption requires that the ideal take into account the two possible vertex correspondences for each of the 2D rotations.

Let $\mathbf{S}$ represent the coordinates of the vertices in a stimulus image, and $\mathbf{T} = \{\mathbf{T}_1(\phi), \mathbf{T}_2(\phi), \ldots, \mathbf{T}_{11}(\phi)\}$ represent the 11 prototype template images, each expressed as a function of rotation angle, ϕ, in the image plane. We will write the probability of obtaining a stimulus $\mathbf{S}$ given that it is the target stimulus as $p_{\text{target}}(\mathbf{S}|\mathbf{O})$. By assumption, $\mathbf{S}$ was generated by adding a fixed level of noise to any one of the prototype template images at any one of the possible 2D rotations. $P_{\text{target}}(\mathbf{S}|\mathbf{O})$ is therefore given by

$$p_{\text{target}}(\mathbf{S}|\mathbf{O}) = \sum_{i=1}^{11} \int_0^{2\pi} \left[\frac{1}{2} p_{\text{target}}(\mathbf{S}|\mathbf{T}_i(\phi)) + \frac{1}{2} p_{\text{target}}(\mathbf{S}'|\mathbf{T}_i(\phi)) \right] p(\mathbf{T}_i(\phi))\, \mathrm{d}\phi, \quad (2)$$

where $p_{\text{target}}(\mathbf{S}|\mathbf{T}_i(\phi))$ is the probability that $\mathbf{S}$ was generated by adding noise to template i at rotation angle ϕ. $\mathbf{S}'$ is the same stimulus as $\mathbf{S}$, but with the order of vertices inverted. $p(\mathbf{T}_i(\phi))$ is the prior probability of the stimulus having been generated from template i at rotation ϕ. This is assumed to be uniform at $p(\mathbf{T}_i(\phi)) = 1/22\pi$. The ideal observer selects the stimulus for which the value of $p_{\text{target}}(\mathbf{S}|\mathbf{O})$ is greatest. Appendix A gives more details of the 2D/2D ideal observer's formulation.

Note that the decision is made by considering possible matches with all the template images. An alternative strategy, which would not perform as well, would be to only use the template which best matched the stimulus, a so-called nearest neighbor classification (Duda & Hart, 1973). Choosing the nearest neighbor does not, in general, produce the maximum of $P(\text{image}|\text{object})$ [see equation (1)].

We have also derived an observer which we call the Learning 2D/2D ideal. This model acquires as a new template after each test trial the average of the two test images presented in the trial. Since both of the test stimuli are noisy versions of the same view of the prototype, their average approximates a projected image of the prototype object. The learning ideal uses the same algorithm for matching as the 2D/2D ideal with the exception that the number of templates it matches against increase with each trial. A more detailed description of this learning ideal is included in Appendix A.

3.4.2.2. The 3D/2D ideal observer. The 3D/2D ideal observer is similar to the 2D/2D ideal, except that instead of matching stimuli against all 2D rotations of the 11 learned template views, it matches against all possible views of the prototype object. An ideal observer which can only detect the 2D vertex positions in a stimulus image, but has a full 3D model of the prototype (the 3D/2D ideal), would estimate the probability that a stimulus is a target by integrating the probabilities of obtaining that stimulus by adding white Gaussian noise to the 2D vertex positions in each of the possible views of the prototype. It takes as its image formation model

$$\mathbf{S} = \mathbf{T}(\phi, \theta, \omega) + \mathbf{N}, \quad (3)$$

where $\mathbf{S}$ is a representation of the 2D vertex positions of a wire in a stimulus image and $\mathbf{T}(\phi, \theta, \omega)$ is the orthographic projection of the 3D vertex positions of a prototype object with θ and ω being the viewing angle and ϕ being the rotation angle around the viewing axis. $\mathbf{N}$ is the noise added to the 2D positions of the vertices in the stimulus image. Let $p_{\text{target}}(\mathbf{S}|\mathbf{O})$ be the probability of obtaining $\mathbf{S}$ as an image of the prototype object after adding the target level of noise. $p_{\text{target}}(\mathbf{S}|\mathbf{O})$ is given by

$$p_{\text{target}}(\mathbf{S}|\mathbf{O}) = \int_0^{\pi}\int_0^{2\pi}\int_0^{2\pi} p_{\text{target}}(\mathbf{S}|\mathbf{T}(\theta, \phi, \omega)) \times p(\theta, \phi, \omega)\,\mathrm{d}\theta\,\mathrm{d}\phi\,\mathrm{d}\omega, \quad (4)$$

where $p_{\text{target}}(\mathbf{S}|\mathbf{T}(\theta, \phi, \omega))$ is the probability of S having been generated by adding the target level of noise to the 2D vertex positions of the prototype object in view, $\mathbf{T}(\phi, \theta, \omega)$. $p(\phi, \theta, \omega)$ is the prior probability of viewing the prototype object from a fixed position on the viewing sphere and at a fixed orientation around the viewing axis.

Unfortunately, the solution of the integral in equation (3) is computationally prohibitive, so we do not actually calculate the performance of this 3D/2D ideal. Instead, we have calculated a nearest-neighbor approximation to the ideal observer in which each stimulus image S is compared with the view of the prototype O to which it is closest (using a Euclidean distance norm). Thus, for each stimulus image S, we search for the orientation (ϕ, θ, ω) from which the 2D projection $\mathbf{T}(\phi, \theta, \omega)$ of the 3D model has the minimal Euclidean distance with S. That is, we seek the global minimum of the function $\|\mathbf{S} - \mathbf{T}(\phi, \theta, \omega)\|$. The stimulus for which this distance is minimized is selected by the ideal as its guess of the target. More details of the 3D/2D ideal are given in Appendix B.

3.4.2.3. The 3D/3D ideal observer. The 3D/3D ideal matches a complete 3D representation of stimulus vertex positions with an internal representation of the prototype's 3D vertex positions. We assumed that the 3D translations and rotations of the input were known, so that the correspondence problem for the matching between the input and the internal representation had been solved. Under these assumptions, we have been able to obtain an analytic solution, in the form of a polynomial equation, for the threshold distractor noise level relative to the target noise level needed to obtain a fixed percentage of correct responses. The derivation is given in Appendix C.

3.4.2.4. Generalized radial basis function (GRBF) models. In addition to comparing human performance to the above ideal observers, we also computed efficiency relative to a GRBF implementation (Poggio & Edelman, 1990). In general, a GRBF model consists of a layer of hidden units each of which computes a Gaussian function of the input vector. The outputs of these units then serve as input to a simple linear associative net, i.e. the weights of the associative net, represented as a matrix, are applied to the output of the hidden units to give the final output of the system. This output could be either a scalar or a vector. In learning, the parameters describing the Gaussian functions for the hidden units (mean vectors and standard deviations) as well as the weights of the linear associative net are modified to optimize the mapping from training inputs to desired outputs. The final system computes the function

$$f(\mathbf{x}) = \sum_{i=1}^{M} w_i G_i(\|\mathbf{x} - \mathbf{t}_i\|; \sigma_i) \quad (5)$$

where $\mathbf{x}$ is the vector input, w_i is either a weight vector or a scalar weight depending on whether the output is a vector or a scalar, $G_i(\bullet)$ is the Gaussian with standard deviation, σ_i and mean vector $\mathbf{t}_i$ associated with the ith hidden unit, and $f(\mathbf{x})$ is the output of the system.

In their GRBF model of object recognition, Poggio *et al.* (Poggio & Edelman, 1990) train a system to recreate from any view of a single object, a prototype view of that object. Their model represents any view of a wire object by a vector of its projected vertex coordinates (in general, they argue that an object could be represented by any appropriate feature coordinates). The system is trained to associate a set of training views of an object with a prototype view. In the comparisons given here, we simulate a specific example of such a system, which has the same number of hidden units as there are training views.* In this case, the means of the Gaussians associated with each hidden unit are simply the template views themselves. The training views were generated from the 11 template views used in the experiment. First, both vertex orderings for the stimulus views were taken into account. Second, for each of the 11 views, a finite number of 2D rotations were included as template views. The number of rotations was selected to be that number at which performance of the model reached an asymptote. More details of the GRBF model we simulate are given in Appendix D.

We note that the GRBF model can be considered to be an approximation to the 3D/2D ideal observer (see Appendix D). It effectively finds the best fitting Gaussian interpolation (from the 11 training views and their 2D rotations for both vertex orderings) to the function mapping the hypersurface defining all possible views of a prototype object to a prototype view. If that function were known, it would capture all the 3D information captured in the 3D/2D ideal.

4. RESULTS

Figure 6 shows the threshold of distractor noise level in centimeters needed for 75% correct classification of target stimuli for each of the four types of prototype objects (the fixed target noise level is indicated by a straight dashed line on the graph). An analysis of variance performed on the results show the following effects: a main effect of object type [$F(3, 6) = 15.46, P < 0.005$], a main effect of view type (old vs new) [$F(1, 2) = 30.76, P < 0.05$], and a significant interaction between object type and view type [$F(3, 6) = 6.80, < 0.025$]. Clearly, subjects' performance

*A simple alternative implementation would be to have a scalar output whose value would indicate how well a test image matches the learned object (Edelman & Poggio, 1990). When the output weights of this model are uniform, this version of GRBF is equivalent to the 2D/2D ideal we describe.

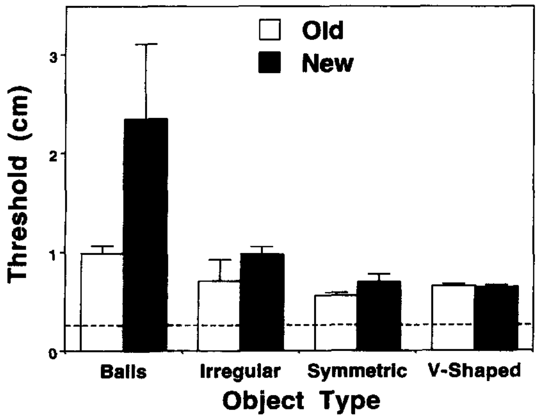

FIGURE 6. Subjects' averaged performance. Thresholds are the mean standard deviations of the Gaussian positional noise distributions added to the distractor vertices, when subjects were 75% correct classifying the targets from the distractors. The targets had a fixed positional noise with the standard deviation of 0.254 cm shown by the dashed line. The target/distractor pairs were generated from the same prototype wire objects. A QUEST procedure was used to adjust the thresholds to keep subjects' performance at 75% correct level. Error bars are ± 1 SD.

improves with increasing regularity of the objects tested and the difference in performance between old views and new views appears to decrease with increasing regularity of the objects tested. A *post hoc* test on view type (new vs old) effect for different objects yielded a significant effects for Balls [$t(2) = 3.31$, $P < 0.05$], a non-significant effect for Irregular [$t(2) = 1.65$], a significant effect for Symmetric [$t(2) = 3.58$, $P < 0.05$], and no effect for V-Shaped [$t(2) = -0.41$].

Thresholds for the 3D/3D ideal observer were computed numerically from the defining polynomial equation derived in Appendix C. Thresholds for the 2D/2D, the 3D/2D ideals, and the GRBF model, which could not be computed analytically, were estimated by running QUEST for 2000 trials. The Learning 2D/2D was simulated using the same sequence of stimuli as subjects were shown. Figure 7(a, b) shows the thresholds for the ideals and the GRBF model for the old and new views, respectively. Performance of the 2D/2D and learning 2D/2D ideals as well as the GRBF model was considerably worse for new views than for old views. This difference does not appear for either the 3D/2D or 3D/3D ideals, as it should not given their definitions. Note that the thresholds do not differ significantly across object type. As one would expect, the 3D/3D observer does best, and the 2D/2D observer does worst. The learning 2D/2D shows a slight improvement over the 2D/2D observer as would be expected since it has more information on which to base its judgments. The performance of the GRBF observer lies between the 3D/2D and the 2D/2D observer.

From the threshold data, we calculated statistical efficiency measures for subject performance relative to the ideals and the GRBF model. Efficiency for this task (see Appendix E for the derivation) is given by the ratio of threshold differences between target and distractor noise variances for an ideal and a subject. Figure 8(a–e) shows subjects' statistical efficiency relative to the ideals and the GRBF model. Efficiencies were computed for each object used in the experiment independently and then averaged within object type to obtain the data shown in the plot. Of particular note in these results is the fact that for the three types of wire objects, subjects were able to match or beat the 2D/2D ideal on new views, as reflected in efficiencies which were greater than or equal to 100%. For new views of the Symmetric and V-Shaped objects, subjects could still beat the Learning 2D/2D ideal and the GRBF model. Figure 8(c) shows the estimates of the 3D/2D efficiencies which are relative to the actual information available in the task. Peak efficiency is for old views and is around 20%. It is interesting to note that 3D/2D efficiencies [Fig. 8(c)] are only about twice the 3D/3D efficiencies [Fig. 8(d)]. This reflects the high efficiency of the 3D/2D ideal relative to the 3D/3D ideal.

Subjects took an average of 3 sec to respond to each trial (the Balls condition took slightly longer). After each experimental session, subjects were asked to describe what they were doing and to draw the prototype objects on which they were just tested. All subjects reported that they tried to imagine the target object as a single 3D object, rather than as a set of images. No subjects mentioned that they had to explicitly decide which end of the objects to match up.

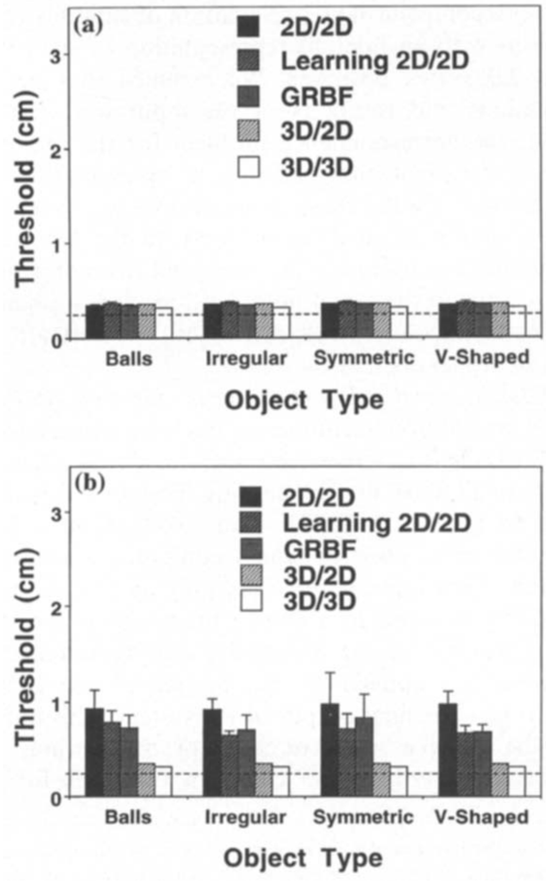

FIGURE 7. (a, b) Mean thresholds over the three objects for the ideal observers and the GRBF model for old and new views, respectively. The error bars show ± 1 SD.

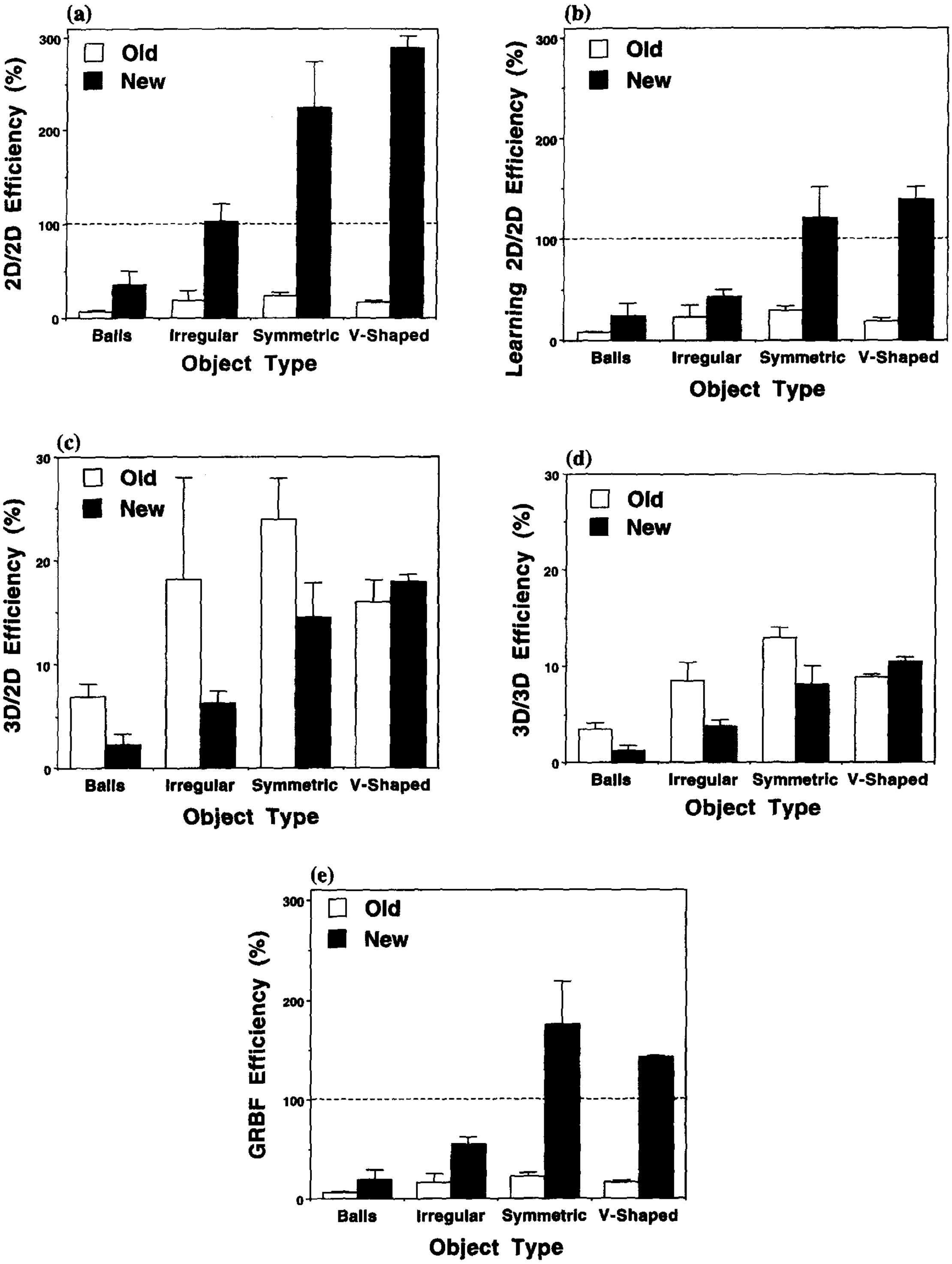

FIGURE 8. Statistical efficiency of human observers for the various object types relative to: (a) the 2D/2D ideal; (b) the learning 2D/2D ideal; (c) the 3D/2D ideal; (d) the 3D/3D ideal; and (e) the GRBF model for the different types of objects. The means were computed by averaging the efficiencies across the three objects for each type. The error bars show ±1 SD.

5. DISCUSSION

5.1. 2D vs 3D Internal Representations

If an ideal model approximates the scheme used by the human visual system, one expects that the efficiency would be less than 100%, because of internal noise; however, the performance pattern should appear the same for the ideal and human observer. One of the main results obtained in the experiment is that for the three types of wire objects used as prototypes, human subjects performed as well as or better than the 2D/2D ideal for new views of the objects. The high efficiency relative to the 2D/2D ideal eliminates a simple 2D template matching strategy as a model for performance in this task. Further, even when the 2D/2D ideal is allowed to learn new templates during the testing phase, human performance still exceeds it for new views of the Symmetric and V-Shaped objects [Fig. 8(b)]. We must, however, consider the possibilities that: (a) although the memory representation may be 2D, the matching process incorporates some 3D constraints, such as the approximation of the view hypersurface learned by GRBF models; and (b) coarse 3D information is stored along with 2D templates and this is matched against coarse 3D information provided in the stimulus images by features such as shading and occlusion. We distinguish this strategy from the strategy of the 3D/2D ideal which has a 3D object model and compares stimulus images with possible views of the object model.

5.1.1. Comparisons with GRBF models

Poggio and Girosi have developed a very powerful theory called Hyper Basis Function approximation (HyperBF) that can be applied to a diverse set of function approximation problems, including recognition (Poggio & Girosi, 1989, 1990). The GRBF models proposed as possible models for recognition (Edelman & Poggio, 1992; Poggio & Edelman, 1990) are special cases of HyperBFs. Depending on the implementation, the performance of a HyperBF model can span the range between our 2D/2D and 3D/2D ideal observers. We restricted our analysis to scalar and vector output GRBFs that assume the number of the centers are equal to that of the learned templates, the standard deviation of the hidden units' Gaussian interpolating functions is fixed, and the mean vectors of the Gaussians are fixed to be the same as the learned template views. Our 2D/2D ideal is equivalent to a scalar-output GRBF model which has as templates all the learned views (with both possible vertex orderings) and their rotations in 2D. The simulations shown in the results were for a stronger, vector-output GRBF model which has as templates not only the 11 training views, but also a densely sampled set of targets corresponding to all 2D rotations of the training views as well as both possible vertex orderings for test stimuli. As shown in Fig. 8(e), subjects' efficiencies relative to this GRBF implementation were greater than 100% for the Symmetric and V-Shaped objects.* The GRBF performance could be pushed closer to that of the 3D/2D ideal if it had more than this set of views, and if allowed the additional power of the HyperBF approximations that are captured by adjustable means, a weighted norm, as well as a polynomial weighting term.

Simple modifications of the specific GRBF model we implemented would not seem to account for our results. It is counterproductive, for instance, to reduce the number or change the means of the hidden units, when the old views—the centers themselves—are the only feedback the model can have. Any such changes would hurt the performance of the model on trials on which old views were tested, which made up half of the test trials. Similarly, although it may improve the model's performance for the new views when the standard deviation of the Gaussian function is changed, it is difficult to achieve this without feedback from the new views. In fact, we have simulated the model's performance for the old views for a few objects, with different Gaussian standard deviations, and found no better standard deviations than the current one used for the old views. One way to save the GRBF model that we simulated would be to propose that new templates were learned by subjects during the testing phase and added as new centers for Gaussian interpolating functions. Although this possibility can never be firmly rejected, we attempted to investigate it by looking at the temporal course of reaction times during the testing phase. One might expect that if learning occurred during the testing phase, reaction times for new view trials relative to old would decrease with time. We divided each session into three equal parts and averaged the reaction times for new and old view trials in each of the three parts of the sessions. Figure 9 shows the results of this analysis. No significant effects of time are apparent in the data. Even if such learning were to occur, one would not expect it to differentially affect performance on Irregular, Symmetric and V-Shaped objects.

Finally, we wish to reiterate that general forms of HyperBF, with appropriate feedback, would be quite a bit more powerful than the GRBF model simulated here, which was not designed specifically for symmetric objects. A HyberBF network can, in fact, learn from examples a good approximation of the hypersurface defining the possible views of an object (Poggio, personal communication). This would provide an implementation of Ullman and Basri's theorem that three 2D views of

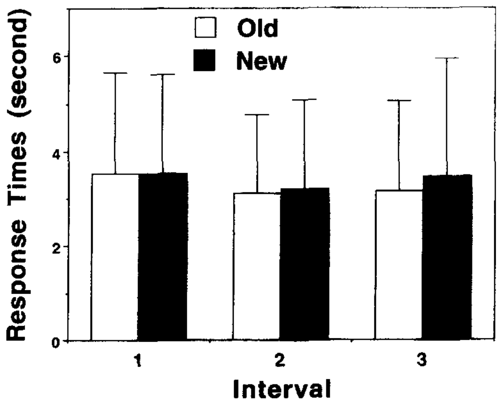

FIGURE 9. Response times for three temporal intervals during testing for each object, pooled over all the objects. Any change of response times for the old views will only indicate a task familiarization effect, but not learning new templates. Therefore, a non significant trend of the change of response times for the new relative to old views suggests that no substantial learning of new templates occurred during the experiment.

*A potential way of improving the performance of the GRBF model (or, for that matter, the 2D/2D ideal) for symmetric objects is suggested by the recent result of Poggio and Vetter (1992) that for each view of a symmetric object, one can generate three novel views (and their 2D rotations). At first glance, it appears that this could explain the improvement in performance for Symmetric and V-Shaped objects. It turns out, however, that the set of training views we used was closed under the operation prescribed by Poggio and Vetter. That is, applying the appropriate operation for generating new views to any of the 11 training views used in the experiment simply resulted in the generation of another view in the set of 11.

a wire object are enough to reconstruct the 3D model of the object itself (Ullman & Basri, 1989). Such a model would provide a good approximation to the 3D/2D ideal we have described.

5.1.2. 3D information in the stimulus

A major assumption underlying the ideal observers and the GRBF model we simulated was that an observer had as input information only the 2D vertex positions of the wire object in a stimulus (the 3D/3D ideal was an exception to this, having as input the 3D vertex positions of a wire object). The stimuli, however, contained coarse depth information which subjects could have used to perform the task. This information took the form of the shading on the arms of the wire objects and the occasional occlusion of one part of an object by another. Subjects could well have employed a mixed strategy of matching both 2D stimulus features and coarse 3D information (e.g. depth ordering from occlusions) with similar information in an internal object model. The 3D information in the stimuli, however, appears to be very coarse, and its incorporation into an ideal observer would not greatly improve the ideal's performance. Previous results did not find any difference between thick wire objects used here and objects of balls connected with lines, where shading on the lines and their occlusion ordering were unavailable (Liu, Kersten & Knill, 1992). The only depth ordering cue available was when one ball occluded another, which happened rarely. Moreover, it seems to us that the use of such information would not, by itself lead to differential performance for new vs old views or for Irregular, Symmetric and V-Shaped objects. As a qualitative test for such biasing effects, we ran simulations to measure how often occlusions occurred in stimulus presentations of the different wire objects used in the experiment. Occlusions occurred approximately as often for all three types of objects, suggesting that the information provided by occlusions is no greater for one type of object than another. Viewpoint dependency effects in recognition may be obtained even when reliable 3D information is provided in the stimuli (Edelman & Bülthoff, 1992). Edelman and Bülthoff showed that the inclusion of stereo information in views of a wire object, while decreasing the absolute error rate, had no effect on the viewpoint dependency of recognition.

5.1.3. Generality of result

Our rejection of the 2D template matching strategy with 2D operations as a model for the current task does not imply that a 2D template matching mechanism is not part of the complete object recognition system. The system may use a hierarchy of strategies, including 2D template matching, in doing object recognition (Edelman, 1991). What strategy human observers use may depend on the particular task as well as the information available for performing the task. Moreover, a sequence of strategies of increasing complexity may be applied until a reliable solution is obtained for a recognition task. For example, a simple 2D template matching strategy may be the first applied to images. When this gives a fairly reliable solution, the solution is accepted, whereas when it doesn't, more complex strategies involving, for example, the reconstruction of 3D representations from image templates may be employed.

5.2. Object Regularity

An effect of object regularity is clearly evident in the data. Not only does subjects' performance improve with increasing regularity of the prototype objects, but the difference in performance between old and new views decreases with increasing regularity. Both results suggest that regularities in the 3D structure of the prototype objects were taken advantage of in the internal representations generated by subjects. Poorest performance was for the Balls condition. As an "object", a collection of balls lacks connectivity and an ordering of feature elements. The lack of ordering is problematic both for setting up correspondences between views and for comparing a new view with existing templates. An example of how regularity in symmetry can improve performance is offered by Poggio and Vetter's (1992) recent theory that a legal 2D view of a 3D object can be determined from only one 2D model view if the object is mirror symmetric [their theory relies on the view combination theory of Ullman and Basri (1991)]. The V-shaped objects had regularities of planarity and colinearity of the two arms. These regularities may facilitate the development of internal representations. Performance may also have been better because colinearity is an important projective invariant; the visual system may be more sensitive to colinearity than other positional changes. Rock and DiVita (1987) have made arguments similar to the above about the role of regularities in discussing a possible hierarchy of object representations. Certainly, such a strategy which utilized object regularities would optimize the efficiency of representations, in terms of memory capacity requirements. Moreover, many of the objects we encounter have strong regularities in shape, such as symmetry, further suggesting that having specialized matching processes for dealing with these would be an efficient allocation of limited resources. It is worth emphasizing that none of the models which we have considered (Fig. 8) can account for the advantage conferred by regularity.

5.3. Viewpoint Dependency

The data indicate that for the balls and symmetric objects, subjects' performance for new views is worse than for the old views, and that it is not dependent on viewpoint for V-shaped objects. Although we did not find a significant difference of view type for irregular objects, this is largely due to one subject's poor performance for one irregular object, which was the very first object on which he was tested. Consistent with the reaction time measurements of previous studies (Bülthoff & Edelman, 1992; Tarr, 1989), we believe that viewpoint dependency for irregular objects may show up if we test more subjects. It seems, therefore, that viewpoint dependency is a function of the object structure. It can diminish when the objects are highly structured, even for the type of subordinate level object classification task used here.

As mentioned above, Rock and DiVita (1987) have argued that the generation of 3D object representations is greatly facilitated by regularities in object structure. Such regularities tap into special purpose representational systems. Irregular objects do not tap into such systems and are therefore represented using an image-based scheme. They argue that only for very regular objects is a 3D representation significantly more efficient than an image-based representation, so that a system designed to maximize efficiency of information storage would be more likely to represent regular objects in 3D than irregular objects. Particularly interesting in this regard is Poggio and Vetter's result, mentioned above, that one view of a bilaterally symmetric object is sufficient to test whether a new picture is a novel view of the same object or not. This suggests that simply adding a flag indicating symmetry to one view of an object is enough to solve the recognition task. We would like to add to this potential explanation of differential viewpoint-dependent effects the fact that extraction of 3D information is greatly facilitated by object regularities, so that the visual system explicitly infers detailed 3D structure only for regular objects. Probabilistic arguments show that a strong inference of object structure can be obtained when it has special properties such as symmetry, even without an explicit indication that the object has such properties. That the visual system should be designed to detect such regularities in 3D structure from images makes a great deal of sense in a world in which objects can be categorized along lines of the different types of regularity (planarity, symmetry, parallelness, colinearity, etc.) which they contain [see Richards and Jepson (1992) for a detailed theoretical discussion of these points]. Given these considerations, as well as the preliminary nature of the results presented here, further research on the effects of object regularity on object recognition should be pursued.

5.4. Usefulness of the Ideal Observer Approach

The ideal observers' performances varied somewhat between different objects indicating that the information of the task was more reliable for some objects than others. This by itself justifies the use of efficiency to obtain an absolute measure of performance, since it factors out effects of information reliability. How did it come into play, however, for the two main stimulus manipulations used in the experiment, viewpoint and object regularity? The 2D/2D ideals and the GRBF model all showed differential performance for old and new views, as did subjects. Whereas, typically, one would be left with a qualitative comparison between these two effects, the use of efficiency allows one to make a quantitative assessment of the relative strength of the viewpoint effect for the ideals and the subjects. The fact that efficiency was higher relative to these ideals for new views than it was for old views allowed us to make an argument against the models which otherwise could not have been made.

In regards to object regularity, Figs 6 and 7 show no significant difference in the performance of any of the ideals or the GRBF model across different object types. This indicates that the differences in information reliability found between different objects were effectively random and not tied to the type of an object. For purposes of comparing absolute performance across different object types, therefore, we could just as well have used raw threshold data. We note, however, that the lack of a significant difference in information reliability between object types could not be known *a priori* and simulation of the ideals was necessary to ascertain that a correct measure of performance be used to assess differences between object types. In more complicated tasks, in which stimuli are created by some heuristic standards, and not randomly as was done here, the possibility of experimenter bias unknowingly affecting the information content of the stimuli increases, making the need for simulating ideal observers for a task even more important.

6. CONCLUSION

To conclude, we have extended the ideal observer paradigm to the study of object perception. That the efficiencies relative to the 3D/2D ideal were on the order of 10–20% suggested that they were comparable with those in other, lower-level perceptual tasks (Barlow & Reeves, 1979), so that experiments like this are reasonable ones to address the problem of 3D object recognition. Based on the comparison between human and ideal observers, we can rule out simple 2D template matching algorithms, for which the 2D/2D ideal observer yields an upper limit on the best possible performance, as mechanisms underlying human performance on the task. The GRBF model we have simulated in this paper yields the best possible performance under the conditions given in (Poggio & Edelman, 1990), but nevertheless cannot account for the performance level in all of our conditions. The increase in efficiency with increases in prototype regularity indicates that special aspects of 3D object structure, such as symmetry and planarity, are taken advantage of in the presentation and recognition of objects. This suggests, in particular, that models of human recognition should include special mechanisms to handle symmetric objects. Finally, the viewpoint dependency of object recognition appears to be dependent on the structure of objects. Metric object recognition can show little or no dependence on viewpoint when the objects are highly regular.

REFERENCES

Anderson, J. A., Silverstein, J. W., Ritz, S. A. & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413–451.

Barlow, H. B. (1980). The absolute efficiency of perceptual decisions. *Philosophical Transactions of the Royal Society of London B, 290,* 71–82.

Barlow, H. B. & Reeves, B. C. (1979). The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research, 19,* 783–793.

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

Biederman, I. & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 1162–1182.

Bienenstock, E. & Von der Malsburg, C. (1987). A neural network for invariant pattern recognition. *Europhysics Letters, 4,* 121–126.

Bülthoff, H. H. & Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science U.S.A., 89,* 60–64.

Duda, R. O. & Hart, P. E. (1973). *Pattern classification and scene analysis.* New York: Wiley.

Edelman, S. (1991). *Features of recognition* (CS91-10). The Weizmann Institute of Science, Israel.

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

Edelman, S. & Poggio, T. (1990). *Bringing the grandmother back into the picture: A memory-based view of object recognition* (A.I. Memo 1181). MIT.

Edelman, S. & 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,* 37–61.

Gerhardstein, P. C. (1992). 3D orientation invariance in objection recognition. Ph.D. thesis, University of Minnesota, Minn.

Huttenlocher, D. P. & Ullman, S. (1987). Object recognition using alignment. In *Proceedings of the International Conference on Computer Vision* (pp. 102–111). London: IEEE.

Kersten, D. (1990). Statistical limits to image understanding. In Blakemore, C. (Ed.), *Vision: Coding and efficiency.* Cambridge: Cambridge University Press.

Koenderink, J. J. & van Doorn, A. J. (1976). The singularities of the visual mapping. *Biological Cybernetics, 24,* 51–59.

Kohonen, T. (1978). *Associative memory: A system theoretic approach.* Berlin: Springer.

Liu, Z., Kersten, D. & Knill, D. C. (1992). Object discrimination for human and ideal observers. *Investigative Ophthalmology and Visual Science (Suppl.), 33,* 825.

Plantinga, H. & Dyer, C. R. (1990). Visibility, occlusion, and the aspect graph. *International Journal of Computer Vision, 5,* 137–160.

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

Poggio, T. & Girosi, F. (1989). *A theory of networks for approximation and learning* (A.I. Memo 1140, C.B.I.P. Paper 31). MIT.

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

Poggio, T. & Vetter, T. (1992). *Recognition and structure from one 2D model view: Observations on prototypes, object classes and symmetries* (A.I. Memo No. 1347, C.B.I.P. Paper No. 49). MIT.

Richards, W. & Jepson, A. (1992). *What makes a good feature?* (C.B.I.P. Paper 72). MIT.

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

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439.

Tarr, M. J. (1989). *Orientation dependence in three-dimensional object recognition.* Ph.D. thesis MIT.

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

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

Ullman, S. & Basri, R. (1989). *Recognition by linear combinations of models* (A.I. Laboratory Technical Report 1152). MIT.

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

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

Watson, A. B. & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics, 33,* 113–120.

Acknowledgements—We thank Irving Biederman, Heinrich Bülthoff, Gordon Legge, Thomas Vetter, and Philippe Schyns, and in particular, Pascal Mamassian for helpful discussions. Two of the reviewers provided extremely helpful suggestions. We thank Tomaso Poggio for many useful comments, and in particular the result that Hyper BFs are capable of implementing the Basri-Ullman theorem. This research was supported by grants NSF BNS-9109514, AFOSR-90-0274, and the Minnesota Supercomputer Institute.

APPENDIX A

The 2D/2D Ideal Observer

The learned templates consist of the 11 images presented to subjects during the training phase plus their 2D rotations in the image plane. We represent these as vector functions of 2D rotational angle ϕ with each function corresponding to one of the learned templates,

$$\mathbf{T} = \{\mathbf{T}_1(\phi), \mathbf{T}_2(\phi), \ldots, \mathbf{T}_{11}(\phi)\}, \tag{A1}$$

where $\mathbf{T}_i(\phi) = \breve{\mathbf{R}}(\phi)\mathbf{T}_i(0)$, $\mathbf{T}_i(0) = [0, 0, X_1, Y_1, \ldots, X_4, Y_4]^T$, and $\breve{\mathbf{R}}(\phi)$ is the 2D rotation matrix. The image formation model for the target stimulus is:

$$\mathbf{S}_{\text{target}} = \mathbf{T}_i(\phi) + \mathbf{N}, \tag{A2}$$

where $\mathbf{N}$ is a 10-dimensional vector of independent random variables, each with a Gaussian distribution of $G(0; \sigma_t)$, and $\mathbf{S} = [0, 0, x_1, y_1, \ldots, x_4, y_4]^T$ represents the coordinates of a stimulus's vertices. As the ideal does not know which end-point of the stimulus to choose, both $\mathbf{S}$ and its end-to-end reversal, $\mathbf{S}' = [0, 0, x_3 - x_4, y_3 - y_4, \ldots, -x_4, -y_4]^T$ must be considered in computing the probability. The ideal selects from the test pair the one with a larger probability as the target. We write $p_{\text{target}}(\mathbf{S}|\mathbf{O})$ as

$$p_{\text{target}}(\mathbf{S}|\mathbf{O}) = \sum_{i=1}^{11} \int_0^{2\pi} \left[\frac{1}{2} p_{\text{target}}(\mathbf{S}|\mathbf{T}_i(\phi)) + \frac{1}{2} p_{\text{target}}(\mathbf{S}'|\mathbf{T}_i(\phi))\right] \times p\,(\mathbf{T}_i(\phi))\, \mathrm{d}\phi, \tag{A3}$$

where $p_{\text{target}}(\mathbf{S}|\mathbf{T}_i(\phi))$ is the probability of having obtained $\mathbf{S}$ by adding noise $\mathbf{N}$ to $\mathbf{T}_i(\phi)$, given by

$$p_{\text{target}}(\mathbf{S}|\mathbf{T}_i(\phi)) = p\,(\mathbf{N} = \mathbf{T}_i(\phi) - \mathbf{S}) = \prod_{j=1}^{4} \frac{1}{\sqrt{2\pi}\,\sigma_t} \times \exp\left[-\frac{(X_j \cos\phi + Y_j \sin\phi - x_j)^2 + (-X_j \sin\phi + Y_j \cos\phi - y_j)^2}{2\sigma_t^2}\right]. \tag{A4}$$

The prior distribution of training views is assumed to be uniform, so that

$$p\,(\mathbf{T}_i(\phi)) = \frac{1}{11}\frac{1}{2\pi} = \frac{1}{22\pi}. \tag{A5}$$

The ideal selects a target from the test pair from each trial. Threshold data were obtained using QUEST on the same stimuli used for the human subjects.

To test the possibility that subjects learned, without feedback, new 2D templates during testing, we have devised a learning 2D/2D ideal. This ideal acquires a new template after each test pair by storing the average of the two stimuli. Recalling that during the experiment, subjects were tested on 240 pairs of images, 120 of which old views, 120 new views, randomly mixed. We, however, have simulated the model for old and new views separately, each with 120 pairs. This is due to the computational limitation and our assumption that adding more noisy old views may little improve the performance of the model for new views, but hurt the performance for the old views, as indeed shown in the figure. Note that, unlike the 2D/2D ideal or the GRBF models, this learning 2D/2D ideal acquires new templates which cannot be obtained by 2D reflection and rotation in the image plane from the stored templates, but truly *new* views. The fact that human subjects can beat even this ideal observer further argues that our result cannot be accounted for by the 2D template matching.

APPENDIX B

The 3D/2D Ideal Observer

The ideal takes as its image formation model

$$\mathbf{S} = \mathbf{T}(\phi, \theta, \omega) + \mathbf{N}, \tag{B1}$$

where $\mathbf{S} = [0, 0, x_1, y_1, \ldots, x_4, y_4]^T$ is the image, $\mathbf{T}(\phi, \theta, \omega) = \tilde{\mathbf{P}}\tilde{\mathbf{R}}(\phi, \theta, \omega)\mathbf{O} = [0, 0, X_1, Y_1, \ldots, X_4, Y_4]^T$ is the orthographic projection of the 3D object $\mathbf{O}$, which is a vector representing the 3D vertex positions of the prototype object, $\tilde{\mathbf{R}}(\phi, \theta, \omega)$ is the viewpoint transformation which rotates the object into a viewpoint-centered coordinate system and $\tilde{\mathbf{P}}$ is the orthographic projection transform, and noise $\mathbf{N}$ is a vector of independent, Gaussian random variables, $\mathbf{N} = [0, 0, n_1, n_2, \ldots, n_8]^T$. The probability of having obtained a stimulus, given that it was created as the target is given by

$$p_{\text{target}}(\mathbf{S} | \mathbf{T}(\phi, \theta, \omega)) = p(\mathbf{N} = \mathbf{T}(\phi, \theta, \omega) - \mathbf{S}). \tag{B2}$$

Integrating this over all possible views gives

$$p_{\text{target}}(\mathbf{S}|\mathbf{O}) = \int_0^{\pi}\int_0^{2\pi}\int_0^{2\pi} p(\mathbf{N} = \mathbf{S} - \mathbf{T}(\theta, \phi, \omega))p(\theta, \phi, \omega)\,\mathrm{d}\theta\,\mathrm{d}\phi\,\mathrm{d}\omega, \tag{B3}$$

where $p(\phi, \theta, \omega)$ is the prior probability of viewing the object from position (θ, ω) at orientation ϕ around the viewing axis. For the experiment this was uniform, so $p(\phi, \theta, \omega) = 1/4\pi^2 \sin\omega$. $p(\mathbf{N} = \mathbf{T}(\phi, \theta, \omega) - \mathbf{S})$ is given by the Gaussian distribution

$$p(\mathbf{N} = \mathbf{T}(\phi, \theta, \omega) - \mathbf{S}) = \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[-\frac{(X_i - x_i)^2 + (Y_i - y_i)^2}{2\sigma_t^2}\right]$$
$$= \frac{1}{(\sqrt{2\pi}\sigma_t)^4} \exp\left[-\frac{\sum_{i=1}^{4}(X_i - x_i)^2 + (Y_i - y_i)^2}{2\sigma_t^2}\right], \tag{B4}$$

so we have for $p_{\text{target}}(\mathbf{S}|\mathbf{O})$

$$p_{\text{target}}(\mathbf{S}|\mathbf{O}) = \int_0^{\pi}\int_0^{2\pi}\int_0^{2\pi} \frac{1}{4\pi^2}\sin\omega \frac{1}{(\sqrt{2\pi}\sigma_t)^4}$$
$$\times \exp\left[-\frac{\sum_{i=1}^{4}(X_i - x_i)^2 + (Y_i - y_i)^2}{2\sigma_t^2}\right]\mathrm{d}\theta\,\mathrm{d}\phi\,\mathrm{d}\omega. \tag{B5}$$

Unfortunately, evaluation of the integral in equation (B5) is computationally prohibitive, so we have chosen to approximate the 3D/2D ideal using a nearest-neighbor model. In this model, we use as a measure of fit between $\mathbf{S}$ and $\mathbf{O}$ the Euclidean distance between the $\mathbf{S}$ and the view of $\mathbf{O}$ which minimizes this distance. Searching for the orientation (ϕ, θ, ω) that minimizes the Euclidean distance between $\mathbf{S}$ and the 2D projection $\hat{\mathbf{T}}(\phi, \theta, \omega)$ of the 3D object $\mathbf{O}$. $\hat{\mathbf{T}}(\phi, \theta, \omega)$ is given by

$$\hat{\mathbf{T}}(\phi, \theta, \omega) = \arg[\min\|\mathbf{T}(\phi, \theta, \omega) - \mathbf{S}\|], \tag{B6}$$

where

$$\|\mathbf{T}(\phi, \theta, \omega) - \mathbf{S}\| = \sum_{i=1}^{4}[(X_i - x_i)^2 + (Y_i - y_i)^2]. \tag{B7}$$

Note that this 3D/2D ideal yields the same performance for old and new views.

APPENDIX C

The 3D/3D Ideal Observer

We assume that the observer has an internal representation of the 3D vertex positions of the prototype and receives as input a specification of the 3D coordinates of the wire vertices in the stimulus images. We further assume that the observer knows the viewing position, so that it can match stimulus objects to prototypes in world coordinates. Let us represent the prototype model as a vector of 3D vertex coordinates

$$\mathbf{O} = [0, 0, 0, X_1, Y_1, Z_1, \ldots, X_4, Y_4, Z_4]^T, \tag{C1}$$

where we have taken one endpoint of the wire to be the origin. The target stimulus is generated by adding noise to the prototype object

$$\mathbf{S}_{\text{target}} = \mathbf{O} + \mathbf{N}, \tag{C2}$$

where $\mathbf{S} = [0, 0, 0, x_1, y_1, z_1, \ldots, x_4, y_4, z_4]^T$ represents the vertex coordinates of a stimulus object and $\mathbf{N} = [0, 0, 0, n_1, n_2, \ldots, n_{12}]^T$ is a vector of independent random variables, each with a Gaussian distribution of G(0; σ_t). By not adding noise to the endpoint of the wire, we can always treat the position of the endpoint of a stimulus as the origin and represent all vertex positions relative to this point. Doing this makes the formulation of all of the ideal observers translation invariant.

We represent the probability of obtaining a stimulus, $\mathbf{S}$, given that it is the target, as $p_{\text{target}}(\mathbf{S}|\mathbf{O})$. In any given trial the ideal observer selects the stimulus which has the larger value of $p_{\text{target}}(\mathbf{S}|\mathbf{O})$. $p_{\text{target}}(\mathbf{S}|\mathbf{O})$ is given by

$$p_{\text{target}}(\mathbf{S}|\mathbf{O}) = p(\mathbf{N} = \mathbf{O} - \mathbf{S})$$

$$p_{\text{target}}(\mathbf{S}|\mathbf{O}) = \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[-\frac{(X_i - x_i)^2 + (Y_i - y_i)^2 + (Z_i - z_i)^2}{2\sigma_t^2}\right]$$

$$p_{\text{target}}(\mathbf{S}|\mathbf{O}) = \frac{1}{(\sqrt{2\pi}\sigma_t)^4}$$
$$\times \exp\left[-\frac{\sum_{i=1}^{4}\{(X_i - x_i)^2 + (Y_i - y_i)^2 + (Z_i - z_i)^2\}}{2\sigma_t^2}\right]. \tag{C3}$$

From equation (C3), we see that the ideal observer can use as a decision variable the Euclidean distance between the stimulus vectors and the prototype vector. The observer should select as the target, the stimulus with the smallest distance. Since we can derive the distribution of this distance for both target and distractor stimuli, we can derive the threshold level of distractor noise needed to reach a given performance level (i.e. percentage correct).

It is easy to show that the squared Euclidean distance, d, between an n-dimensional vector and the same vector disturbed by white Gaussian noise G(0; σ) has a χ^2 distribution

$$p_d(d; n) = \frac{d^{(n/2-1)}}{2^{n/2}\sigma^n\Gamma(n/2)} \exp\left[\frac{-d}{2\sigma^2}\right]. \tag{C4}$$

Given a fixed σ_t, we would like to find the standard deviation σ_d for the distractor noise which would give a threshold probability of obtaining a squared Euclidean distance for the target stimulus greater than that obtained for the distractor stimulus, i.e. $P(d_t > d_d) = P_{\text{threshold}}$ (in our case, $P_{\text{threshold}} = 0.75$). Equivalently, we want to find σ_d so that $P(d_t - d_d < 0) = P_{\text{threshold}}$. To do that we need to derive the probability density function of $\Delta d = d_t - d_d$, $p_{\Delta d}(\Delta d; n)$.

Since d_t and d_d are independent, we can write $p_{\Delta d}(\Delta d; n)$ as

$$p_{\Delta d}(\Delta d; n) = \begin{cases} \int_0^{\infty} p_{d_t}(x; n)p_{d_d}(x - \Delta d; n)\,\mathrm{d}x; & (\Delta d < 0) \\ \int_{\Delta d}^{\infty} p_{d_t}(x; n)p_{d_d}(x - \Delta d; n)\,\mathrm{d}x; & (\Delta d \geqslant 0) \end{cases} \tag{C5}$$

where the region of integration differs for the two ranges of Δd because of the singularities in $p_{d_t}(d_t; n)$ and $p_d\mathrm{d}(d_d; n)$ at 0. Since we are only concerned with the range $\Delta d < 0$, we need only evaluate the first integral. Expanding (C5), we obtain

$$p_{\Delta d}(\Delta d; n) = \frac{1}{2^n[\Gamma(n/2)]^2(\sigma_t\sigma_d)^n}\int_0^{\infty} x^{n/2-1}\exp\left[\frac{-x}{2\sigma_t^2}\right](x - \Delta d)^{n/2-1}$$
$$\times \exp\left[\frac{-(x - \Delta d)}{2\sigma_d^2}\right]\mathrm{d}x$$
$$= \frac{\exp[\Delta d/2\sigma_d^2]}{2^n[\Gamma(n/2)]^2(\sigma_t\sigma_d)^n}\int_0^{\infty} x^{n/2-1}(x - \Delta d)^{n/2-1}$$

$$\times \exp\left[-x\left(\frac{1}{2\sigma_{\mathrm{t}}^2}+\frac{1}{2\sigma_{\mathrm{d}}^2}\right)\right]\mathrm{d}x$$

$$= \frac{\exp[\Delta d/2\sigma_{\mathrm{d}}^2]}{2^n[\Gamma(n/2)]^2(\sigma_{\mathrm{t}}\sigma_{\mathrm{d}})^n}\int_0^\infty \sum_{m=0}^{n/2-1} C^m_{n/2-1} x^{n-m-2}(-\Delta d)^m$$

$$\times \exp\left[-x\left(\frac{1}{2\sigma_{\mathrm{t}}^2}+\frac{1}{2\sigma_{\mathrm{d}}^2}\right)\right]\mathrm{d}x,$$

where $C^p_q = p!/q!(p-q)!$. Making the substitution, $z = x(1/2\sigma_{\mathrm{t}}^2 + 1/2\sigma_{\mathrm{d}}^2)$, and assuming n is even (true in our case), we obtain

$$p_{\Delta d}(\Delta d; n) = \frac{\exp[\Delta d/2\sigma_{\mathrm{d}}^2]}{2^n[\Gamma(n/2)]^2(\sigma_{\mathrm{t}}\sigma_{\mathrm{d}})^n}\left(\frac{1}{2\sigma_{\mathrm{t}}^2}+\frac{1}{2\sigma_{\mathrm{d}}^2}\right)^{-n+m+1}$$

$$\times \sum_{m=0}^{n/2-1}\left\{C^m_{n/2-1}(-\Delta d)^m\int_0^\infty z^{n-m-2}\exp(-z)\,\mathrm{d}z\right\}.$$

Noting that

$$\int_0^\infty z^p \exp(-z)\,\mathrm{d}z = p!,$$

for positive integer p, we obtain finally

$$p_{\Delta d}(\Delta d; n) = \frac{\exp[\Delta d/2\sigma_{\mathrm{d}}^2]}{2^n[\Gamma(n/2)]^2(\sigma_{\mathrm{t}}\sigma_{\mathrm{d}})^n}\left(\frac{1}{2\sigma_{\mathrm{t}}^2}+\frac{1}{2\sigma_{\mathrm{d}}^2}\right)^{-n+m+1}$$

$$\times \sum_{m=0}^{n/2-1}\{C^m_{n/2-1}(-\Delta d)^m(n-m-2)!\} \quad \text{(C6)}$$

for $\Delta d < 0$. The ideal observer's percentage correct is given by the probability that $\Delta d < 0$; thus, the distractor noise level σ_{d} needed to achieve a fixed threshold performance level satisfies

$$P_{\text{threshold}} = \int_{-\infty}^0 p_{\Delta d}(\Delta d; n)\,\mathrm{d}\Delta d. \quad \text{(C7)}$$

Expanding (C7) gives

$$P_{\text{threshold}} = \int_{-\infty}^0 \frac{\exp[\Delta d/2\sigma_{\mathrm{d}}^2]}{2^n[\Gamma(n/2)]^2(\sigma_{\mathrm{t}}\sigma_{\mathrm{d}})^n}\left(\frac{1}{2\sigma_{\mathrm{t}}^2}+\frac{1}{2\sigma_{\mathrm{d}}^2}\right)^{-n+m+1}$$

$$\times \sum_{m=0}^{n/2-1}\{C^m_{n/2-1}(-\Delta d)^m(n-m-2)!\}\mathrm{d}\Delta d.$$

Making the substitution, $u = -\Delta d/2\sigma_{\mathrm{d}}^2$, and after algebraic manipulation of the terms involving σ_{t} and σ_{d}, we have

$$P_{\text{threshold}} = \frac{(\sigma_{\mathrm{d}}^2/\sigma_{\mathrm{t}}^2)^n}{[\Gamma(n/2)]^2(1+\sigma_{\mathrm{d}}^2/\sigma_{\mathrm{t}}^2)^n}\sum_{m=0}^{n/2-1}\left\{C^m_{n/2-1}(n-m-2)!\right.$$

$$\left.\times\left(1+\frac{\sigma_{\mathrm{d}}^2}{\sigma_{\mathrm{t}}^2}\right)^{1+m}\int_{-\infty}^0 \exp(-u)u^m\,\mathrm{d}\Delta d\right\}$$

$$= \frac{(\sigma_{\mathrm{d}}^2/\sigma_{\mathrm{t}}^2)^n}{[\Gamma(n/2)]^2(1+\sigma_{\mathrm{d}}^2/\sigma_{\mathrm{t}}^2)^n}$$

$$\times \sum_{m=0}^{n/2-1}\left\{C^m_{n/2-1}(n-m-2)!\left(1+\frac{\sigma_{\mathrm{d}}^2}{\sigma_{\mathrm{t}}^2}\right)^{1+m} m!\right\}.$$

Noting that $\Gamma(n/2) = (n/2-1)!$ when n is even (as in our case), we can simplify to

$$P_{\text{threshold}} = \frac{(\sigma_{\mathrm{d}}^2/\sigma_{\mathrm{t}}^2)^n}{(n/2-1)!(1+\sigma_{\mathrm{d}}^2/\sigma_{\mathrm{t}}^2)^n}$$

$$\times \sum_{m=0}^{n/2-1}\left\{\frac{(n-2-m)!}{n/2-1-m)!}\left(1+\frac{\sigma_{\mathrm{d}}^2}{\sigma_{\mathrm{t}}^2}\right)^{1+m}\right\}. \quad \text{(C8)}$$

Equation (C8) indicates that the ideal observer's performance is determined by the ratio of distractor and target noise variances, $\sigma_{\mathrm{d}}^2/\sigma_{\mathrm{t}}^2$, a fact of which we will take advantage in our derivation of an efficiency measure for the task (Appendix E). For the experiment here, $n = 12$ (4 vertices with 3 coordinates each), numerical simulation of the 3D/3D ideal gave a threshold ratio corresponding to $\sigma_{\mathrm{d}}^2/\sigma_{\mathrm{t}}^2 = 1.490$ for a 75% correct threshold level. We verified this result by running QUEST for an ideal which based its decisions on the Euclidean distance between stimuli and a prototype, obtaining a threshold corresponding to $\sigma_{\mathrm{d}}^2/\sigma_{\mathrm{t}}^2 = 1.498$ after 2000 trials.

APPENDIX D

The Generalized Radial Basis Function (GRBF) Model

Poggio and Edelman proposed a specific form of HyperBF model for object recognition, termed a generalized radial basis function model (GRBF) (Poggio & Girosi, 1990). The model learns a mapping from a number of 2D images of an object to a prototype image of the object. An image is represented as a vector, $[x_1, y_1, x_2, y_2, \ldots, x_N, y_N]^T$, of the 2D coordinates of N feature points in the images. For the example of wire objects, as used also in Poggio and Edelman's paper, the feature points are the projected vertices of the wires. The mapping is performed by a three layer network with an input layer, a layer of hidden units, and an output layer. The input layer codes the vector representing an input image and the output layer codes the output representing the reconstructed prototype vector. The hidden layer of units applies K Gaussian basis functions to the input vector producing a K-dimensional vector of outputs which is transformed by a linear weight matrix to produce the reconstructed prototype vector. Formally, the system performs the following mapping

$$\mathbf{O}(\mathbf{S}) = \sum_{i=1}^K \mathbf{w}_i G_i(\|\mathbf{S}-\mathbf{T}_i\|; \sigma_i), \quad \text{(D1)}$$

where $\mathbf{O}(\mathbf{S})$ is the reconstructed prototype, $\mathbf{w}_i$ is the weight vector associated with hidden unit i, and $G_i(\|\mathbf{S}-\mathbf{T}_i\|; \sigma_i)$ is the Gaussian basis function associated with hidden unit i. The Gaussian has mean vector $\mathbf{T}_i$ and standard deviation σ_i. The parameters of the system, weight vectors, the means and standard deviations of the Gaussian basis functions, are adaptively learned when the system is trained to associate a set of training views of an object with a prototype view $\mathbf{P}$. The transformation given in (D1) is applied to novel images of objects to generate an estimate of the associated prototype vector. The Euclidean distance between this reconstruction and the learned prototype vector, $\|\mathbf{O}(\mathbf{S}) - \mathbf{P}\|$, gives a measure of goodness of fit between the input image and the learned object. If the distance exceeds some threshold, the image is rejected as a candidate view of the learned object.

In our experiment, direct application of the GRBF model would require that the model simply choose which of two stimuli result in a smaller Euclidean distance between the corresponding reconstructed vector and the learned prototype. In our simulations, we set the number of hidden units equal to the number of images on which subjects were trained. Technically, this would only give 11 hidden units; however, we assume subjects do not know the vertex ordering, and that they have available all 2D rotations of the training views. The first assumption means we must include stimulus views corresponding to both possible vertex orderings. The second assumption implies that there are potentially an infinite number of training views. We uniformly sample the rotation space finely enough (M samples) to achieve maximum performance of the model (see below), giving $11 \times M$ training views and $K = 11 \times M$ hidden units. In the case that the number of hidden units equals the number of training views, the weight vectors of the GRBF model can all be set to point in the same direction as the prototype vector, i.e. $\mathbf{w}_i = c_i\mathbf{P}$, where $\mathbf{P}$ is the prototype vector and c_i is a scalar factor. Under these conditions, the selection of the prototype vector plays no role in the performance of the model as can be seen by the following expansion of the Euclidean distance metric

$$\|\mathbf{O}(\mathbf{S}) - \mathbf{P}\| = \left\|\sum_{i=1}^K \mathbf{w}_i G_i(\|\mathbf{S}-\mathbf{T}_i\|; \sigma_i) - \mathbf{P}\right\|$$

$$= \left\|\sum_{i=1}^K c_i\mathbf{P} G_i(\|\mathbf{S}-\mathbf{T}_i\|; \sigma_i) - \mathbf{P}\right\|$$

$$= \|\mathbf{P}\|\left(\sum_{i=1}^K c_i G_i(\|\mathbf{S}-\mathbf{T}_i\|; \sigma_i) - 1\right). \quad \text{(D2)}$$

That is, the metric applied to each of two stimulus images will be scaled by the same constant $\|\mathbf{P}\|$ and $\mathbf{P}$ appears nowhere else in the expansion. Even under this constraint, the solution to which parameters to select for the hidden units' Gaussian basis functions based on the set of training views is an underdetermined problem. We set the mean vectors to be equivalent to the training vectors and the standard deviations to be constant at the standard deviation used for the target stimuli. Having fixed the parameters of the Gaussian basis functions, we are left with only the scalar weights c_i to be determined to find a unique mapping which takes each of the training views to the prototype view. These need not be learned adaptively, as they are given as the solution to the set of $22 \times M$ independent linear equations

$$\sum_{i=1}^{K} c_i G_i(\|\mathbf{S}-\mathbf{T}_i\|;\sigma_i) = 1. \tag{D3}$$

We note the similarity between this version of GRBF and the 2D/2D ideal. The difference lies in the fact that the GRBF model computes a weighted sum of Gaussians centered on the training views; whereas, the ideal computes a straight sum. In the simulations of the GRBF model, we computed the c_i independently for each object learned and ran the model on the experimental task using QUEST for 2000 trials to obtain thresholds.

We noted that the rotation space was sampled at a fine enough density to obtain peak model performance. We ran simulations of the GRBF model for a range of sampling densities, from 10 rotations (giving a total of 110 training views) to 100 rotations (giving a total of 1100 training views). Performance of the model asymptotes at about 40 rotations which is well before the peak sampling of 100 rotations is reached. For the data shown in the paper, the density of 100 rotations was used.

It should be noted that a HyperBF model is a general mathematical technique for function approximation (Poggio & Girosi, 1989). Its flexibility and generality makes it useful in a wide range of applications, e.g. in non-parametric statistics. We cannot and do not intend to argue against the general mathematical method. We simply are comparing human performance on the experimental task with a specific implementation of GRBF, keeping as true to the spirit of the model presented in (Poggio & Girosi, 1990) as possible. The model can be generalized quite a bit by allowing different representations for input and output, using radial basis functions other than circularly symmetric Gaussians, varying the number of hidden units, allowing the parameters of the basis functions to be dynamically altered during learning, adding a polynomial term to the Gaussians, and so on. We further simulated the model by varying the size of the standard deviation of the Gaussian functions. This size of the standard deviation of all the Gaussians was kept the same, and a threshold was obtained with the QUEST procedure. By varying the size of the standard deviations, the best threshold value was obtained for each object. No substantial improvement of the model was found, however.

APPENDIX E

Efficiency for Object Discrimination

Efficiency for signal detection and discrimination tasks is often defined as the squared ratio of d's for the ideal and human observer:

$$E = \frac{d'^{(\mathrm{H})2}}{d'^{(\mathrm{I})2}}. \tag{E1}$$

Efficiency defined in this way can often be interpreted as the ratio of the number of samples required by ideal and human observers to perform a task at a given level of performance. This interpretation allows one to discuss the amount of information available in a stimulus effectively used by human observers to perform a task. If the efficiency is one, then one can say that humans are using all the information available in the stimulus to perform the task. If it is less than one, then one can say that human observers are performing as if they were only using a fraction of the samples available (with the fraction given by E).

An alternative interpretation of E is that it gives the ratio of the variance in the signal noise which the observer "sees" and the signal noise which the ideal sees. To understand this, consider the following simple detection task. Subjects are asked to detect, in the presence of noise, a signal, X. For a given signal strength, x, the ideal observer's d' is given by $d'^{(\mathrm{I})} = x/\sigma$. The human observer, when modeled as being limited only by the noise of the stimulus and some independently added internal noise, has a d' which is given by $d'^{(\mathrm{H})} = x/\sqrt{\sigma^2+\sigma_{\mathrm{H}}^2}$, where $\sqrt{\sigma^2+\sigma_{\mathrm{H}}^2}$ is the standard deviation of the stimulus noise plus the internal noise. We therefore have, for the efficiency,

$$E = \frac{d'^{(\mathrm{H})2}}{d'^{(\mathrm{I})2}} = \frac{\sigma^2}{\sigma^2+\sigma_{\mathrm{H}}^2}. \tag{E2}$$

The latter interpretation of efficiency is probably more appropriate for the recognition discrimination task we have performed. Since only a few sample points define the signal (number of vertices multiplying two (or three)), one can hardly imagine that human observers are limited by the number of samples they can process in a trial. Intuitively, one thinks of their performance being limited by added uncertainty at the positions of the vertices (either in memory or in the signals). We would therefore like to define efficiency in the second way described above. We will show that efficiency defined in this way for the object discrimination task is given by

$$E = \frac{\Delta\sigma^{(\mathrm{I})2}}{\Delta\sigma^{(\mathrm{O})2}}, \tag{E3}$$

where $\Delta\sigma^{(\mathrm{I})2} = \sigma_{\mathrm{d}}^{(\mathrm{I})2} - \sigma_{\mathrm{t}}^{(\mathrm{I})2}$ is the threshold difference in noise variances of distractor and target stimuli needed for the ideal observer to obtain a given percentage correct, and $\Delta\sigma^{(\mathrm{O})2} = \sigma_{\mathrm{d}}^{(\mathrm{O})2} - \sigma_{\mathrm{t}}^{(\mathrm{O})2}$ is the threshold difference needed for human observers to obtain the same level of performance.

As in the standard analysis of efficiency, we will treat human observers as being ideal but having internal noise added to the input signals (in this case, the position of the vertices). We have shown in Appendix C that the percentage correct for the ideal observer is determined by the ratio of standard deviations of target and distractor noise,

$$P_{\text{threshold}} = f\left(\frac{\sigma_{\mathrm{t}}}{\sigma_{\mathrm{d}}}\right), \tag{E4}$$

where σ_{t}, is the standard deviation of the target noise and σ_{d} is the standard deviation of the distractor noise. For a fixed threshold percentage correct, we therefore have

$$\frac{\sigma_{\mathrm{t}}^{(\mathrm{I})2}}{\sigma_{\mathrm{d}}^{(\mathrm{I})2}} = \frac{\sigma_{\mathrm{t}}^{(\mathrm{O})2}+\sigma_{\mathrm{H}}^2}{\sigma_{\mathrm{d}}^{(\mathrm{O})2}+\sigma_{\mathrm{H}}^2}, \tag{E5}$$

where $\sigma_{\mathrm{t}}^{(\mathrm{O})2}$ and $\sigma_{\mathrm{d}}^{(\mathrm{O})2}$ are the target and distractor noise levels needed for a human observer to obtain the same threshold percentage correct as the ideal, and σ_{H}^2 is the internal noise added to the vertex positions. Noting that the target noise level is held constant throughout the experiment, $\sigma_{\mathrm{t}}^{(\mathrm{I})2} = \sigma_{\mathrm{t}}^{(\mathrm{O})2} = \sigma_{\mathrm{t}}^2$, and solving for σ_{H}^2, we obtain

$$\sigma_{\mathrm{H}}^2 = \frac{\sigma_{\mathrm{t}}^2(\sigma_{\mathrm{d}}^{(\mathrm{O})2} - \sigma_{\mathrm{d}}^{(\mathrm{I})2})}{\sigma_{\mathrm{d}}^{(\mathrm{I})2} - \sigma_{\mathrm{t}}^2}. \tag{E6}$$

For efficiency, we will use the ratio of the noise variances in the distractor stimulus seen by the ideal and the human observers. We choose to compute efficiency using the distractor as the signal, since the variance of the target is constant for ideal and human observers

$$E = \frac{\sigma_{\mathrm{d}}^{(\mathrm{I})2}}{\sigma_{\mathrm{d}}^{(\mathrm{O})2}+\sigma_{\mathrm{H}}^2}. \tag{E7}$$

For $\sigma_{\mathrm{d}}^{(\mathrm{O})2}+\sigma_{\mathrm{H}}^2$, we have

$$\sigma_{\mathrm{d}}^{(\mathrm{O})2}+\sigma_{\mathrm{H}}^2 = \frac{\sigma_{\mathrm{d}}^{(\mathrm{I})2}(\sigma_{\mathrm{d}}^{(\mathrm{O})2} - \sigma_{\mathrm{t}}^2)}{\sigma_{\mathrm{d}}^{(\mathrm{I})2} - \sigma_{\mathrm{t}}^2}. \tag{E8}$$

Substituting into the expression for E, we obtain

$$E = \frac{\sigma_{\mathrm{d}}^{(\mathrm{I})2}(\sigma_{\mathrm{d}}^{(\mathrm{I})2} - \sigma_{\mathrm{t}}^2)}{\sigma_{\mathrm{d}}^{(\mathrm{I})2}(\sigma_{\mathrm{d}}^{(\mathrm{O})2} - \sigma_{\mathrm{t}}^2)} = \frac{(\sigma_{\mathrm{d}}^{(\mathrm{I})2} - \sigma_{\mathrm{t}}^2)}{(\sigma_{\mathrm{d}}^{(\mathrm{O})2} - \sigma_{\mathrm{t}}^2)} = \frac{\Delta\sigma^{(\mathrm{I})2}}{\Delta\sigma^{(\mathrm{O})2}} \tag{E9}$$

the desired result.