PERGAMON

Vision Research 38 (1998) 2335–2350

Vision Research

The viewpoint complexity of an object-recognition task

Bosco S. Tjan [a,*], Gordon E. Legge [b]

[a] *Max-Planck-Institut für biologische Kybernetik, Spemannstraße 38, D-72076 Tübingen, Germany*
[b] *Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN 55455, USA*

Received 7 January 1997; received in revised form 31 July 1997

Abstract

There is an ongoing debate about the nature of perceptual representation in human object recognition. Resolution of this debate has been hampered by the lack of a metric for assessing the representational requirements of a recognition task. To recognize a member of a given set of 3-D objects, how much detail must the objects' representations contain in order to achieve a specific accuracy criterion? From the performance of an ideal observer, we derived a quantity called the view complexity (VX) to measure the required granularity of representation. VX is an intrinsic property of the object-recognition task, taking into account both the object ensemble and the type of decision required of an observer. It does not depend on the visual representation or processing used by the observer. VX can be interpreted as the number of randomly selected 2-D images needed to represent the decision boundaries in the image space of a 3-D object-recognition task. A low VX means the task is inherently more viewpoint invariant and a high VX means it is inherently more viewpoint dependent. By measuring the VX of recognition tasks with different object sets, we show that the current confusion about the nature of human perceptual representation is partly due to a failure in distinguishing between human visual processing and the properties of a task and its stimuli. We find general correspondence between the VX of a recognition task and the published human data on viewpoint dependence. Exceptions in this relationship motivated us to propose the view–rate hypothesis: human visual performance is limited by the equivalent number of 2-D image views that can be processed per unit time. © 1998 Elsevier Science Ltd. All rights reserved.

Keywords: Object recognition; Perceptual representation; Viewpoint effects; Ideal observer

1. Introduction

How does the human visual system represent its knowledge about objects so that they can be recognized from different views? There are two contending theories, one relies on viewer-centered representations, and the other, object-centered representations [1,2]. Results from human experiments are mixed: some experiments show that recognition performance is viewpoint dependent, while others show that it is viewpoint invariant. What is missing in the attempts to interpret these results is an explicit account of the viewpoint influence due to factors external to the human subjects. Two key external factors, which we shall jointly refer to as the 'task', are the stimuli used and the judgment required of an observer. The purpose of this paper is to evaluate how task characteristics (especially the shapes of the 3-D stimuli) affect viewpoint dependency in object recognition. Conclusions about perceptual representation cannot be drawn without first quantifying the performance constraints imposed by the task. To this end, we shall develop a technique to measure the degree of viewpoint dependency inherent in the task, independent of the observer. Some of the published results on the viewpoint dependence of human performance will be interpreted using our measurements.

Once the task conditions are specified, the nature of perceptual representation may have an impact on performance. Marr and Nishihara [3] suggested that objects can be represented in terms of volumetric primitives. An object-centered representation stores one model per object and is viewpoint independent. To acquire 3-D primitives from a 2-D input image, Biederman [4] proposed that volumetric primitives (geons) with viewpoint-invariant 2-D features can be used as building blocks for a 3-D representation. Ullman [5] suggested a process of feature alignment to match the

* Corresponding author. Tel.: +49 7071 601610; fax: +49 7071 601616; e-mail: bosco.tjan@tuebingen.mpg.de.

0042-6989/98/$19.00 © 1998 Elsevier Science Ltd. All rights reserved.
PII: S0042-6989(97)00255-1

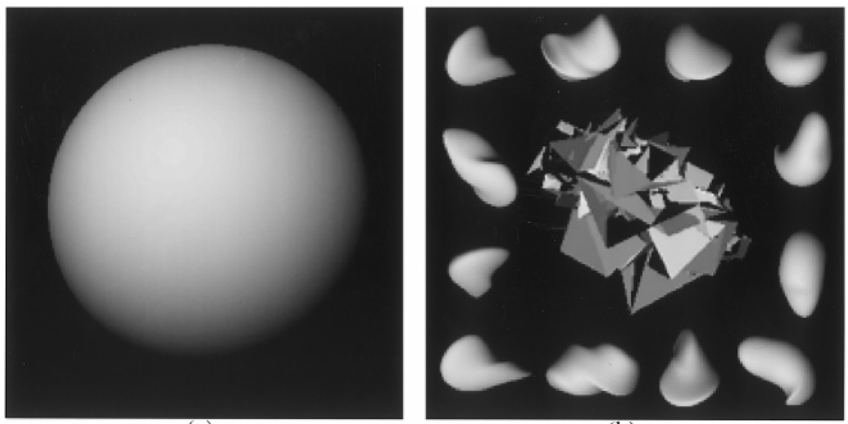

Fig. 1. Regardless of perceptual representation, viewpoint invariance is expected when an object has few discriminable views (a) or contains distinctive global characteristics that set it apart from other objects in its context (b).

2-D input to a 3-D representation. Alternatively, an object can be represented by a collection of 'views' constructed from various 2-D features. Since it is generally impossible to represent every view of an object equally well, recognition performance relying on such a viewer-centered representation is expected to be viewpoint dependent. Viewer-centered representations such as those based on 2-D image components [6], 2-D coordinates of extracted image features [7,8] and linear combinations of feature coordinates from neighboring views [9], were shown to be promising in machine-vision applications.

To test which type of representation is used in the human visual system, a typical experiment consists of, first training or priming a subject with objects from some views and then measuring the subject's recognition performance (naming, memory recall, same–different judgment, *etc*.) as a function of the depth-rotation angle between the test views and the learned/primed views. A small or no viewpoint effect is taken to support theories of object-centered representation, while a large viewpoint effect is taken to favor the viewer-centered theory. While authors have sometimes commented on the nature of the stimulus objects used in these studies, no quantitative measures have been available to equate or rank order sets of stimuli in terms of their inherent viewpoint characteristics within the task.

Using line drawings of novel objects composed of geons, Biederman and colleagues showed that human object recognition is viewpoint independent [10–12]. Other studies, including those using objects such as novel 2-D characters [13], 3-D wire and 'amoebae' [14–16], synthetic animals [17], doll-like figures [18], elongated multi-geon shapes [19]), man-made objects [20] and faces [21–23], show that recognition is viewpoint dependent.

Most researchers looked to properties of the visual system in explaining these discrepant findings. Biederman and Gerhardstein [12] suggested that the viewpoint effect is sometimes caused by visual processes other than those responsible for normal object recognition, while Tarr and Bülthoff [1] maintained that both viewpoint-invariant and viewpoint-dependent mechanisms are integral parts of the visual system. An exception is the proposal by Liu [20] that since the input image must be somehow matched to the perceptual representation of an object, the viewpoint from which the input image is taken can have an effect on this matching process, regardless of the type of perceptual representation.

Still missing from this debate is an objective measure of the viewpoint effect solely due to the task (*i.e*. the combination of the object ensemble and the judgment required of an observer). An object may have very few distinguishable views [Fig. 1(a)], or it may contain global characteristics that set it apart from the rest of the objects in its context [Fig. 1(b)]. In either case, we expect such an object to exhibit substantial viewpoint invariance, regardless of perceptual representation. These examples make it clear that it is risky to assume that human performance provides direct evidence about perceptual representation. Before attributing performance to the nature of an observer's perceptual representation, we argue that the viewpoint characteristics of the object set must be taken into account.

We shall introduce a measure called 'view complexity' (VX) that summarizes the viewpoint properties of a set of objects in a recognition task. The VX of a recognition task is measured independently from the representation used by a visual system. This is done by studying the performance of an ideal observer. By definition, an ideal observer yields performance that is limited only by the informational constraints associated with the task. VX measures how detailed a set of objects must be represented in order to ensure optimal-

ity in recognition accuracy. We consider it to be an objective measure of the degree of viewpoint invariance inherent in an object-recognition task.

2. Theory

Our objective is to characterize the viewpoint properties of a task in a way that does not depend on the mechanisms or algorithms of any particular visual system. Consider a 3-D object-recognition task in which an observer, who knows all views of all objects in the task, is asked to identify an object from any viewpoint. A direct characterization of this 3-D task with a continuum, of views is difficult. Instead, we begin by considering a finite version of this task, the *f*-task, in which objects can, be seen only from a finite number of randomly chosen viewpoints, which are known to the observer. The basic idea is that as the number of views allowed in an *f*-task grows, an *f*-task will approach a continuous 3-D task. We want to express the 'VX' of a 3-D object-recognition task in terms of the number of views needed for an *f*-task to approximate the 3-D task.

2.1. Ideal observer

The input space of either a 3-D task or an *f*-task is taken f to be an array of luminance values (*i.e.* an image). This space defines the initial form of representation that any visual system must deal with in order to perform the task. The decision boundaries in the input space represent how a decision algorithm classifies every possible input in the task. The optimal decision boundaries are those that result from an optimal decision rule, which yields the highest average accuracy at any given signal (contrast) and noise (to be defined later) level. In order to achieve optimality, a decision algorithm, called an ideal observer, must respond in a way that maximizes the *a posteriori* probability (*i.e.* the probability that the response is correct given the input image) [24]. Because of the optimality requirement, the configuration of the optimal decision boundaries is completely determined by the task and the assumed input noise model.

If we assume gaussian luminance noise as the generic noise model for the input (image) space (for reasons to be discussed later), then the decision rule for an *f*-task that maximizes the *a posteriori* probability, is to say that the target in the noisy input is object *i*, if *i* maximizes the following expression [25]:

$$L'(i) = \sum_j \exp\left(-\frac{1}{2\sigma^2}\|R - T_{ij}\|^2 \right) p(T_{ij}) \quad (1)$$

Here, σ is the standard deviation of the noise, R is input image, T_{ij} is the view j of the object i and $p(T_{ij})$ is the prior probability of T_{ij} (see the Appendix for computing this decision rule when the number of object views exceeds 70000 per task). The summation sign in this formula means that all possible views are taken into account in the decision. It distinguishes this optimal decision rule from a nearest-neighbor classifier, which bases the decision on only the best-matched view (*i.e.* the summation is replaced by a 'maximum-of' operator) and is therefore suboptimal [24].

2.2. Scope of analysis

Eq. (1) is a direct mathematical consequence, given (1) an *f*-task, (2) the requirement of maximal accuracy at any signal-to-noise ratio and (3) the gaussian noise assumption. No consideration of visual mechanisms, perceptual representation or processing speed is used in the formulation of Eq. (1). Any decision algorithm that is not equivalent to Eq. (1), violates at least one of the three given, leading to a set of different decision boundaries. We must stress that Eq. (1) is not a model of human object recognition, nor does it suggest a practical implementation of any machine-vision system. Its requirement of comparing all views (translation, rotation and scaling) of each object before making a decision means that it will be too slow and require too large a memory space for any practical vision, system, biological or otherwise. What it allows us to do, however, is to characterize the decision space due to the task. Consideration of algorithmic speed and speed–accuracy trade-off are irrelevant for this purpose.

At a given accuracy criterion, the optimal decision boundaries of a task determine the threshold signal-to-noise ratio (SNR, defined in the Appendix) required of an ideal observer to meet the criterion. The ideal threshold SNR therefore represents a gross summary of the optimal decision boundaries of a task. We chose to report threshold SNR (as opposed to threshold contrast) because it is independent of the noise levels [25].

2.3. Gaussian noise assumption

The reasons for assuming gaussian luminance noise are as follows. First, some form of noise is needed to characterize the performance of an ideal observer. Without any noise, the ideal observer's accuracy for any task will be 100% (unless distinct objects in the task can produce identical images). Second, our stimuli are presented as an array of luminance values. Without assuming extraction of higher-order features, which is a property of a visual system, we must accept each pixel in the image as a distinct 'feature'; to perturb each feature, pixel noise is therefore needed. Third, by convention and mathematical convenience (large-number theorem, maximum entropy consideration, *etc.*), independent gaussian perturbations are often assumed whenever an 'unknown' noise process is required to represent the uncertainty of a feature value.

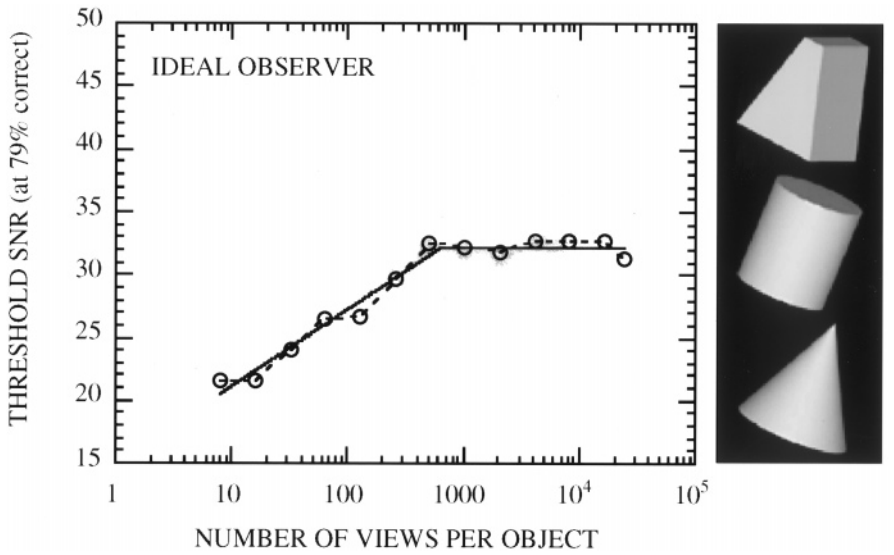

Fig. 2. Typical performance of an ideal observer based on Eq. (1). Threshold SNR for recognizing three simple geometric objects increases initially but then reaches a plateau as the number of allowable views in a 'finite' task, or f-task, increases. The solid line represents a bi-linear model used to describe the data.

2.4. View complexity defined

Fig. 2 shows the threshold SNR at 79%-correct criterion as a function of the number of views allowed per object in an f-task. The f-task involved three objects (details of the simulation will be discussed later). As the number of object views called 'task views' in an f-task grows, the ideal threshold SNR increases as a result of growing viewpoint uncertainty and confusion between views of different objects. When the number of task views becomes very large, ideal performance for the f-task approaches an asymptote. We postulate that this asymptote is stable and equal to the ideal performance for the continuous 3-D task, which we cannot measure directly. We can summarize the ideal observer's behavior with a bi-linear fit to the curve of threshold SNR *versus* log number of task views per object, with the slope of the right branch set to zero. For most of the object sets tested, this bi-linear model provides a good fit to the data ($R > 0.95$). We define the VX of a 3-D recognition task to be the number of views per object used in the corresponding f-task where the breakpoint of the bi-linear fit occurs. In addition, we define the total-VX of a task to be the task's VX (number of views per object) times the number of objects in the task[1]. Note that both VX and total-VX as defined are measurements for the entire object set, and not for any particular object in the set. The VX of a single object in the context of other objects will be defined in Section 4.5.

[1] This definition of total-VX assumes that each participating object has the same prior probability.

2.5. Interpretation of view complexity

VX is operationally defined in terms of the ideal threshold SNR and the viewpoint uncertainty of a task (number of views allowed in an f-task). Assume that (1) as the ideal threshold SNR of a series of f-tasks approaches a plateau value, the corresponding decision boundaries stabilize and (2) the observed plateau actually represents the asymptotic value of threshold SNR as the number of views allowed in an f-task approaches infinity. Then, the VX of a continuous 3-D task is the finite number of random views per object needed to approximate the optimal performance level associated with the 3-D task. In other words, a corresponding f-task with VX number of views per object has the similar decision boundaries in the input space and imposes similar fundamental limits on achievable performance as the continuous 3-D task.

If one 3-D task has a higher VX than another, then its decision boundaries are of higher complexity. A single view of an object becomes less representative, generalization across views is poorer, and more views are needed to define the decision boundaries. This is illustrated in Fig. 3. In other words, a task with a higher VX is inherently less viewpoint invariant.

2.6. View complexity and statistical learning theory

The notion of VX is related to two important concepts in statistical learning theory, that of convergence rate and Vapnik–Chervonenkis (VC) dimension [26]. An algorithm can learn to classify an input as a view of an object by modifying the parameters of an underlying classifier based on training views. Convergence rate

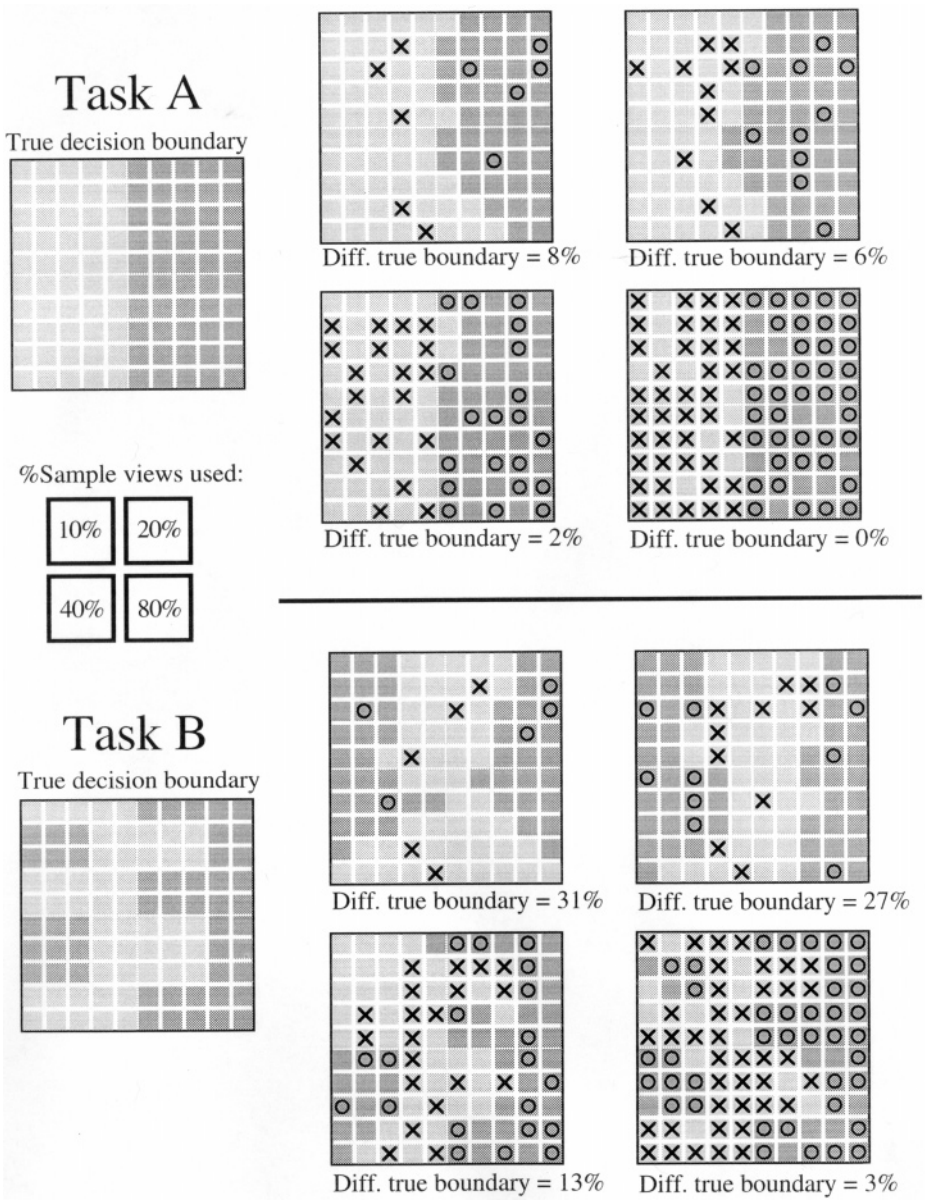

Fig. 3. A simple decision space leads to low VX. Consider two simplified recognition tasks, A and B, in which each view of an object can be described by two feature values and represented by a square in the above depiction of the feature space (left column). The shading of each square represents the object identity of a view. There are two objects in each task, 50 views per object. The decision boundary of task A is a linear partition, while that of task B is more complex. Assuming the feature values are subjected to gaussian noise, the decision boundary for each task can be approximated by a sample of object views ('X's from the 'light-square' object and 'O's from the 'dark-square' object), using the decision rule of Eq. (1). The approximated decision space is depicted by shades of the squares in the middle and right columns. For task A, the decision boundary can be adequately approximated with a smaller number of sample views than for task B. Task A is said to have a lower VX.

relates a learning algorithm's accuracy in classifying new views to the number of training examples it has been given. Theorems in statistical learning theory link convergence rate to the 'power' of the underlying classifier that the learning algorithm uses. This power is expressed in terms of VC dimension. A classifier with a higher VC dimension can establish decision boundaries of a higher complexity than a classifier with a lower VC dimension. However, a high VC dimension generally leads to a low convergence rate. If VC dimension is infinite, the convergence rate becomes undefined.

Note that for a given accuracy criterion, VX measures the number of views needed to approximate the decision boundaries of a task, while convergence rate indicates the number of training views needed by a learning algorithm to learn the task. In theory, we could define the complexity of a task to be the convergence rate of the fastest learning algorithm for that task, working from 2-D images. This, however, presupposes that we know the minimum VC dimension that would be required to classify the objects in image space.

For an arbitrary set of ordinary objects, determining the minimum VC dimension is difficult (see support-vector learning in ref. [26] for an approach to this problem). The operationally defined VX measurement, however, bypasses this need. It provides a practical means of assessing task complexity, albeit without the same analytical rigor. Linking VX to complexity measures in statistical learning theory remains an interesting topic for future research.

In the remainder of this paper, we will describe the measurement of VX for several object-recognition tasks with different object sets. We will discuss the implications of our results on the current debate regarding perceptual representation.

3. General method

3.1. Stimuli

All objects were rendered under orthographic projection from their 3-D models using an SGI Power Indigo 2 graphic workstation with the Open Inventor 3-D graphics library. Lambertian shading with 256 gray levels was used, assuming a point light source at infinity, 21° up and 15° left from the line of sight. The position of the light source was fixed with respect to the observer, and there was no ambient light[2]. Unless otherwise specified, the views of an object used in a task were chosen by randomly and uniformly sampling the surface of a viewing sphere, followed by a random rotation in the image plane. All images were 128×128 pixels in size. The center of the viewing sphere was at the origin of the 3-D object models. While not prevented by Eq. (1), no translation and scaling were applied to the objects, to reduce simulation time. The origin of each object was always placed at the center of the image.

Six types of objects were used in the experiments: geometric objects, pea-like objects, bent-wire objects, mechanical parts, faces and charm bracelets. They are displayed in Figs. 4 and 8 in Sections 4.1 and 4.4, respectively. The geometric objects were selected because each of them is a single 'geon', which, according to the theory of recognition-by-components [4], is a viewpoint-invariant building block for human perceptual representation. The pea-like objects were formed by twisting and bending an elongated and flattened sphere. They are smooth objects that do not have parts. In this respect, they resemble the 'amoebae' objects used in Bülthoff and Edelman [15]. The wire objects are another kind of object frequently used to study human object recognition. Different wire objects were built from the same set of 'geon' parts connected to each other at different angles. They cannot be recognized solely by the identity of their components. The charm bracelet objects are composites formed by combining a geometric object (the charm) with a wire object (the bracelet). Lastly, the two types of complex objects, mechanical parts from a model car's front-end suspension system and human faces, were chosen to demonstrate the generality of the VX method.

The 3-D faces (courtesy of Nikolaus Troje [23]) were acquired from human subjects using a Cyberware 3-D scanner. Because the hair was not digitized, we restricted the viewpoints to only the frontal hemisphere. Since some viewing orientations for faces are unlikely, we also restricted the viewpoints of a face to be sampled from a two-degree-of-freedom space instead of a three-space. Specifically, we required the head's medial axis to have zero tilt with respect to the observer. The axis could have any slant and the face could rotate freely about this axis. This corresponds to viewing a person's face when both the person and the observer are standing upright (therefore, no inverted faces) but not necessarily at the same height or facing each other.

A different static gaussian luminance noise pattern was generated for each trial in the simulation. This was done by first generating a noise pattern of uniform distribution between zero and one with 8192 levels of quantization. Then, a look-up table of an inverse cumulative gaussian distribution was used to convert the uniform noise pattern into a gaussian noise pattern. The size of the noise pattern was the same as the image.

3.2. Simulation

There is no analytical solution for obtaining the threshold SNR as a function of number of task views (Fig. 2) from the ideal decision rule (Eq. (1)). Numerical simulation was therefore used. Time and disk space limited us to 100000 views per recognition task (see the Appendix). All tasks involved three objects. Ideal thresholds were obtained for f-tasks in which the number of task views allowed per object equaled 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384 and 24576.[3] It is possible that within a task, different objects may require different degrees of representation, and hence a different VX; however, simultaneous estimation of distinct VX for different objects turned out to be theoretically difficult. For the simulations reported here, all objects in an f-task

[2] Light source properties form part of the task specification, which can affect the VX of a task. For example, when the light is directly behind an object, only the silhouettes of the object are visible. The VX of such a task can be different from the VX of a similar task when the light source is in front of the objects.

[3] We could, in theory, handle 33000 views per object for a three-object task. The figure of 24576 views per object was chosen because at an early stage of development, four objects (instead of three) were used for the pilot tasks.

had the same number of views. A single VX (in views-per-object) is reported for each task. The VX so measured is likely to be determined by the most complex object in the task.[4] This restriction was relaxed in Section 4.5, in which we measured the VX of just one object in a task. This is done by varying the number of views for just one object, while adequately representing the rest of the objects with a constant number of views.

When the number of task views in an f-task grows large, it becomes prohibitively time-consuming to test the ideal observer for all task views (the same limitation is often encountered in human experiments, where it is impossible to test a subject with every view of a 3-D object). We dealt with this problem using Monte Carlo simulations. Among all task views allowed in an f-task, four sets of eight views per object were randomly selected as the test views.[5] Each test set was presented to the ideal observer in a block of 600 trials, each trial with a different noise pattern. The ideal observer did not 'know' the particular sampling used to select the test views. It assumed that any of the task views could be a test view and considered all of the task views before making a response. For each of the four test sets and for each number of task views, a threshold SNR for the 79%-correct criterion was determined by running a binary search over a range of SNR values. The average simulation time needed to create one data plot (such as one panel in Fig. 4) was about 4 days on an SGI Power Indigo 2 computer.

In Section 4.2, we will show that our sampling of the test views and task views was adequate and led to stable results.

3.3. Data analysis

Simulation results from each test set were fitted with a bi-linear model. The general form of the model is:

$$\mathrm{SNR} = \begin{cases} \gamma(\log(v) - \beta) + \alpha, & \text{if } \log(v) \le \beta \\ \alpha, & \text{otherwise} \end{cases} \quad (2)$$

where v is the number of views allowed per object in an f-task, α is the asymptotic level of SNR, β is the log number of views per object at which the breakpoint occurs and γ is the slope of the curve before the breakpoint. The 'General Curve-Fit' feature of KaleidaGraph 3.0.2 on a Macintosh computer, which implemented the Levenberg–Marquardt least-square method for non-linear curve fit [27], was used to estimate the free parameters α, β and γ.

For each task, the curve-fit was applied separately to the threshold data obtained from each of the four test sets described in Section 3.2. The mean and standard error of the four β values were calculated. $10^{\mathrm{mean}(\beta)}$ is reported as the estimated value of VX for the task. Since the additive standard error (SE) in β corresponds to a multiplicative error in VX, the error in VX is reported in the form of $*/10^{\mathrm{SE}(\beta)}$, which is taken to mean that the estimation of VX varies between $\mathrm{VX}*10^{\mathrm{SE}(\beta)}$ and $\mathrm{VX}/10^{\mathrm{SE}(\beta)}$ with $P = 68\%$. If the threshold SNR did not asymptote within 24576 views per object (our simulation limits), VX was reported as '>25000.'

When analysis of variance (ANOVA) was used to compare measurements across conditions, $P < 0.05$ was adopted as the criterion for an effect to be significant. Because of the way VX was estimated, an analysis on VX was always performed on the values of log VX, and the F-ratio and mean-square error are reported in the same log space.

4. Results

4.1. Experiment 1: VX for different sets of objects

Fig. 4 shows simulation results for five sets of three objects. For each set, a graph shows the relationship between the ideal threshold SNR and the number of views used per object. For each set of objects, separate curves are shown for the four sets of test views.

The set of geometric objects has a low VX, of 700 views per object. In contrast, both of the pea- and wire-object sets have very high VX, exceeding our measurement limit of 25000 views per object. This indicates that the geometric object set is inherently more viewpoint invariant than either the 'pea' or the 'wire' sets.

The mechanical-part ensemble has a VX of 920 views per object, not significantly different from that of the geometric objects [single-factor ANOVA (on log VX); $F(1,6) = 0.2665$, $\mathrm{MS_e} = 0.099$, $P = 0.62$]. Although these mechanical parts are composed of multiple geometric parts, their VX is not significantly greater than the single-part objects.

When the faces were restricted to a hemisphere (no viewing from the back) with only two degrees-of-freedom in orientation (no image-plane rotation), the VX for the task was found to be 3400 views per object, significantly greater than that of the geometric objects [$F(1,6) = 10.401$, $\mathrm{MS_e} = 0.0897$, $P < 0.05$]. When their orientations were unrestricted in the hemisphere, the VX exceeded our measurement limit of 25000. This means the face set is inherently more viewpoint dependent than the geometric objects and the mechanical parts, but if these faces are only viewed upright, they are not as viewpoint dependent as the 'peas' or 'wires'.

4.2. Experiment 2: reliability of VX estimation

To bring the simulation time down to a practical level, we had to sample the viewpoint space. As de-

[4] A complex object in a task does not always have a complex shape. It is complex in the sense that it has to be represented to a greater degree before it can be discriminated from other objects.

[5] When the number of task views per object was <32, different sets of task views were used for different sets of (eight-per-object) test views.

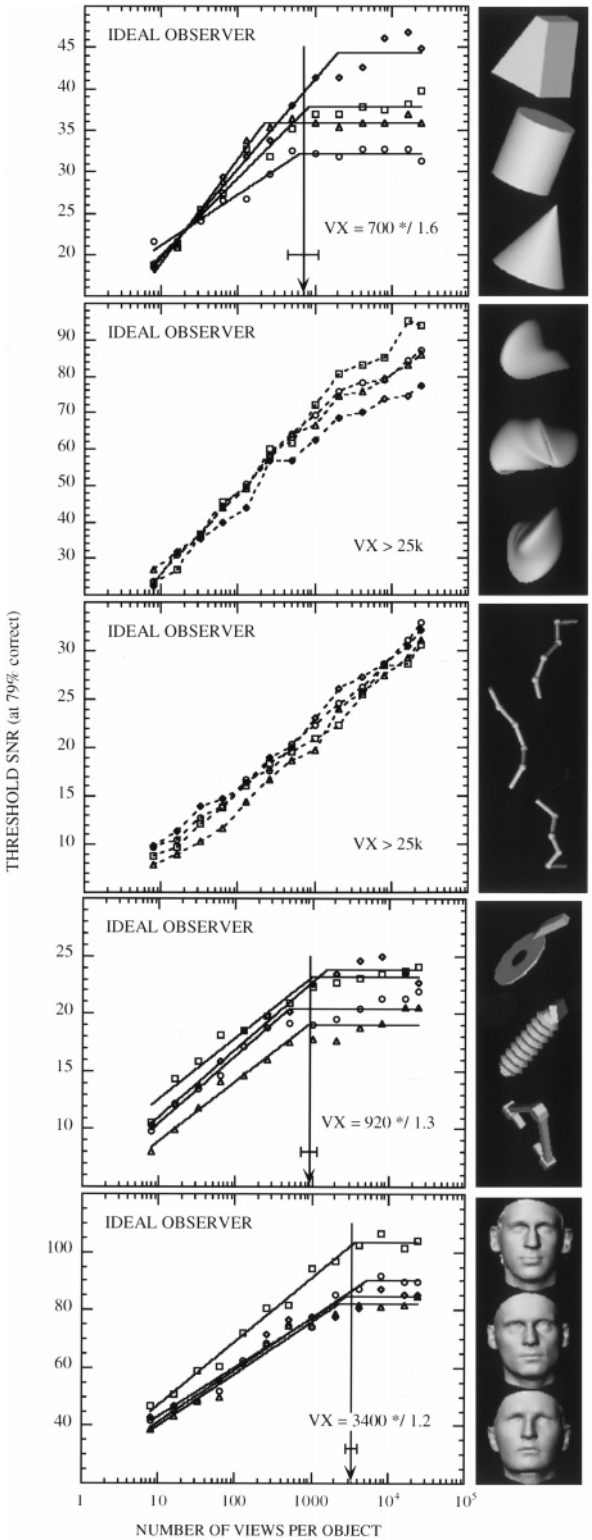

Fig. 4. The VX of five different sets of 3-D objects: simple geometric objects, 'pea' objects, 'wire' objects, mechanical parts, and faces. Each data plot contains four sets of ideal threshold SNR, obtained from four Monte Carlo runs, as a function of the number of views allowed in an f-task. Viewpoints for all objects, except the faces, were uniformly and randomly sampled from the viewing sphere with three degrees-of-freedom. Views for the faces were sampled from the frontal hemisphere and with the restriction that the medial axis had zero tilt (faces were viewed upright). The medial axis can have non zero slant (views from below or above) and the faces were free to rotate about the axis (views from left or right).

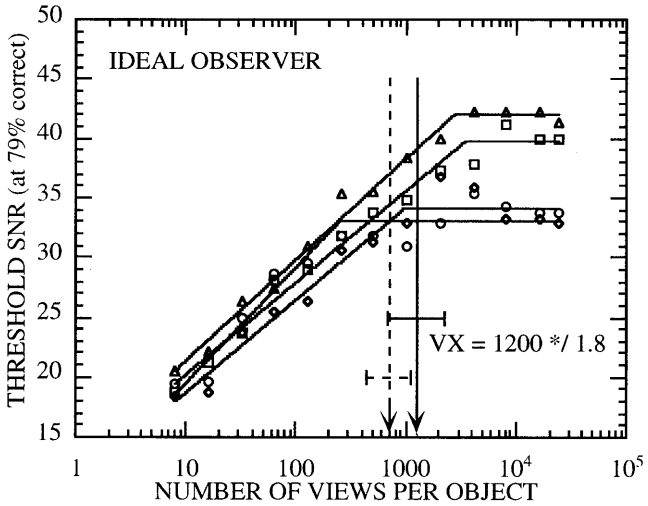

Fig. 5. Repeated VX estimate for the same set of geometric objects studied in Section 4.1, using different samples of task and test views. The VX obtained in Section 4.1 is shown with two a dashed arrow. There is no significant difference between the two estimates.

scribed in Section 3, the task views for each 3-D object were randomly selected from a continuum of views. Even at the maximum of 24576 views per object, the sample represents a set of measure zero from the continuum. In addition, for each Monte Carlo run, the test views (views presented to the ideal observer) were small subsamples from the set of task views. Is the VX measurement sensitive to the specific sets of task or test views? To test this, we re-measured the VX of the geometric-object set using a different random-number generator seed, thus changing both the task- and test-view samples used in the simulation. Fig. 5 shows that the VX estimated with the new samples was 1200 views per object, with an SE of */1.8. A single-factor ANOVA revealed that the difference is not significant [$F(1,6) = 0.5783$, $MS_e = 0.2124$, $P = 0.48$].

If VX really does specify the number of views that are required for a complete representation of the decision boundaries for a given accuracy criterion, then an observer who uses Eq. (1) as the decision rule and stores only VX number of views per object should attain an equal level of performance, regardless of whether the test views are part of the task (*i.e.* stored) views. Within measurement error, this prediction is supported by the data in Fig. 6.

For the set of geometric objects, a two-way ANOVA of mixed designs[6] indicated that threshold SNR varied

[6] The test-view type (known *vs.* unknown) was a between-subjects independent variable. The number of stored views (x-axis of Fig. 6) was a within-subjects independent variable. Threshold SNR was the dependent variable. Data with the number of stored views between eight and 32 were excluded because the observer failed to reach the threshold accuracy criterion in the unknown test view condition.

significantly with both the number of stored views [$F(9,54) = 3.9823$, $MS_e = 45.556$, $P < 0.01$] and the test view type (known *vs.* unknown views) [$F(1,6) = 20.079$, $MS_e = 254.59$, $P < 0.01$]. The ANOVA also revealed a strong interaction between the two [$F(9,54) = 11.431$, $MS_e = 45.556$, $P < 0.01$]. Planned comparisons confirmed that the threshold SNR was greater for the unknown than for the known views when the number stored views was less than VX [$F(1,6) = 38.328$, $MS_e = 193.69$, $P < 0.01$] and there was no significant difference between the two [$F(1,6) = 2.5868$, $MS_e = 186.27$, $P = 0.16$] when the number of stored views exceeded VX.

4.3. Experiment 3: effect of performance criterion on VX

If an ideal observer was to perform at chance level, it could do so without looking at the stimulus; the task becomes viewpoint invariant with a VX of zero. We expect VX to increase as the performance criterion of a task increases. This is consistent with the intuition that the higher the performance level, the more details of the objects need to be represented. Fig. 7 validates this intuition. For the geometric-object ensemble, VX increased from 50 views per object at a criterion of 40% correct to 2100 views per object at a criterion of 95% correct. Over the range of 40–95%, there is an approximately linear relationship between log VX and performance criterion.

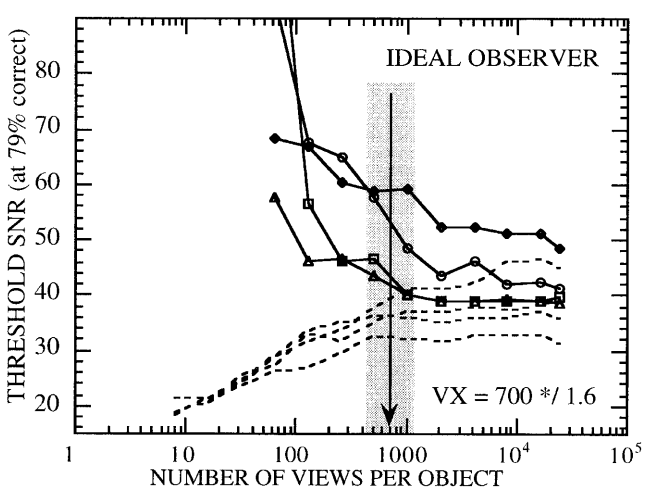

Fig. 6. The threshold SNR as a function of the number of stored views is plotted for the set of geometric objects. The measurements are based on test views that were sampled either from the stored views (the lower dashed curves, 'known views') or from the objects without restrictions (the upper solid curves, 'unknown views'). VX of the task is indicated by a vertical line, with the gray region indicating one SE. When the number of stored views per object exceeded VX, the observer's performance obtained with the unknown views is compatible with that obtained with the known views.

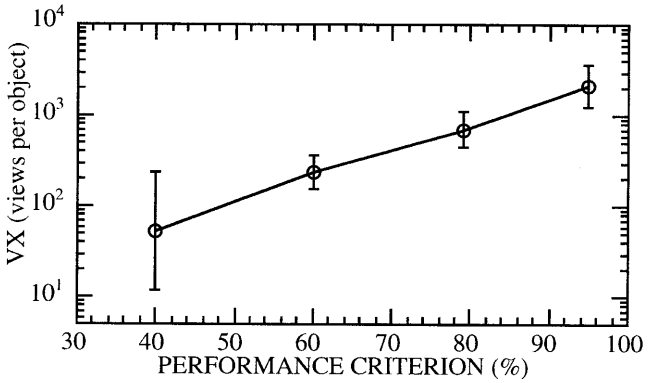

Fig. 7. Effect of performance criterion on the VX of the geometric object ensemble. All conditions are identical to those used in Section 4.1, with the exception that the performance level at which the threshold SNR was obtained ranged from 40 to 95%. VX increases exponentially as the performance criterion increases.

4.4. Experiment 4: VX of compound objects

Some objects have unique and identifiable parts. How does the VX of these compound objects depend on the VX of their parts? In Section 4.1, we showed that the geometric object set had a low VX (700 views per object) while the wire object set had a high VX (> 25000). We formed a set of 'charm bracelets' by combining each geometric object with a different wire object. The VX of this set turned out to be high. Fig. 8 shows that the VX of the charm bracelets is $>$ 25000 views per object, similar to that of the wire objects (Fig. 4). This is because both the low-VX geon part and the high-VX wire part convey object-identity information. To maximize task accuracy, both sources of information must be represented and utilized. As a result, the parts with the highest VX dictate the VX of the compound objects.

4.5. Experiment 5: VX of a single object in a context

We have, so far, defined VX for an object-recognition task, which in our usage of the term, includes the entire stimulus set. Similarly, we can define VX for an individual object in the context of the other objects. This is done by, again, measuring the change in threshold SNR as a function of f-task size and fitting the result with Eq. (2). However, from f-task to f-task, only the number of views allowed for the object of interest changes. The number of views used for the other objects in the task is fixed at a constant greater than or equal to the VX of the task.

We used the geometric-object set in this experiment. Except for the object of interest, the number of views used for all other objects in the task was set at 24576, over 30 times greater than the VX of the geometric-object task. We chose such a large number to avoid any potential sampling artifacts, as well as to allow room

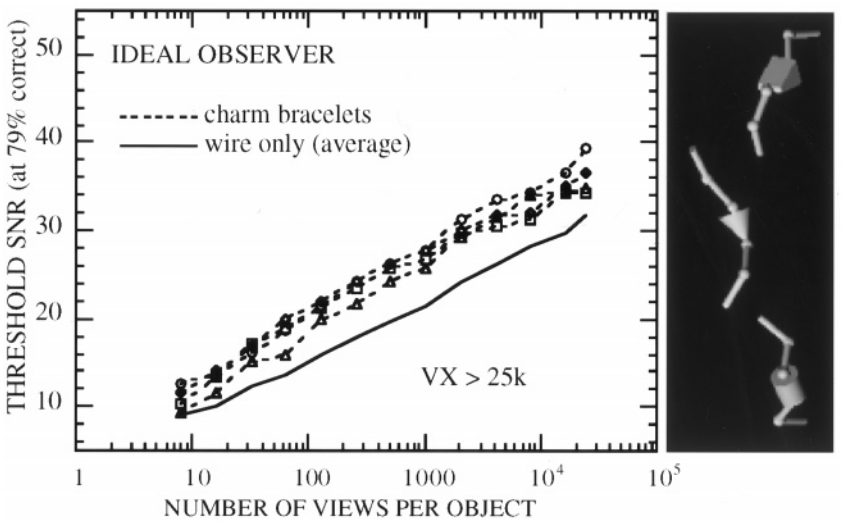

Fig. 8. VX of the charm bracelets made from combining a unique geometric object with a unique wire object. The object set has a high VX (>25000). For comparison, the average threshold curve of the wire-object set, derived from Fig. 4, is plotted as the solid curve.

for the VX of the task to change when we replaced the object of interest.

Fig. 9 shows four measurements of single-object VX. The first three measured the single-object VX of a cone, a pyramid and a wire object in the context of a wedge and a cylinder. The fourth one measured the single-object VX of a cone in the context of a wedge and a pyramid. The VX of each task is also plotted as a comparison. In the context of wedge and cylinder, the cone has a VX of 65 views, significantly lower than that of the pyramid, which has a VX of 2600 views [$F(1,6) = 74.832$, $MS_e = 0.0685$, $P < 0.01$]. This shows that a cone is inherently more viewpoint invariant than a pyramid in this context, consistent with intuition. The wire object, however, has a very low VX (12 views) in this context, which is not significantly different from that of the cone [$F(1,6) = 1.1773$, $MS_e = 0.8802$, $P = 0.32$]. Moreover, replacing the cone with the wire object in this context did not significantly increase the VX of the task [$F(1,6) = 0.7688$, $MS_e = 0.3135$, $P = 0.41$].

Recall that the task involving a set of three wire objects has a very high VX (>25000, Section 4.1). We reason that the single-object VX of each wire in that context must also be very high. The massive difference in VX for the wire object between contexts shows that the extent to which an object must be represented to support recognition is strongly context-sensitive. A less dramatic demonstration of this context effect is to compare the VX of a cone in two slightly different contexts: one with a wedge and a cylinder, the other with a wedge and a pyramid (left-most and right-most conditions in Fig. 9). By a slight change of the context, the single-object VX of the cone increased significantly from 65 to 420 views [$F(1,6) = 8.423$, $MS_e = 0.1560$, $P < 0.05$]. In other words, the cone became less viewpoint invariant when the cylinder in the context was replaced with a pyramid.

These results show that the viewpoint properties of specific objects are highly context dependent. Clearly, context should be considered in studies of viewpoint effects.

5. Discussion

5.1. Limitations on VX measurements

We mention two limitations on the VX measurements. The first relates to the noise model. We stated that VX measures the complexity of the decision space of an object-recognition task independent of the observer. To define the decision space, one needs a representation for the stimuli at the input level. For a typical psychophysical task, this is the computer screen, modeled as an array of luminance values corrupted by gaussian noise. If human performance is inconsistent with VX measurements (*e.g.* human subjects may show less viewpoint dependence in a higher-VX task), we can conclude that factors outside the task, and therefore, within the visual system, are at work. While this achieves an important separation of task factors from visual processing factors, this type of result often does not tell us what those visual-processing factors are. One way to continue the pursuit is to construct VX measurements that are based on certain assumptions of visual processing. Suppose that human vision is limited by gaussian luminance (or luminance-equivalent) noise at early stages (in the photo-receptors, ganglion cells, *etc.*), but by spatial–structural noise at some later stage (IT cortex, *etc.*). Then, the observer-independent VX

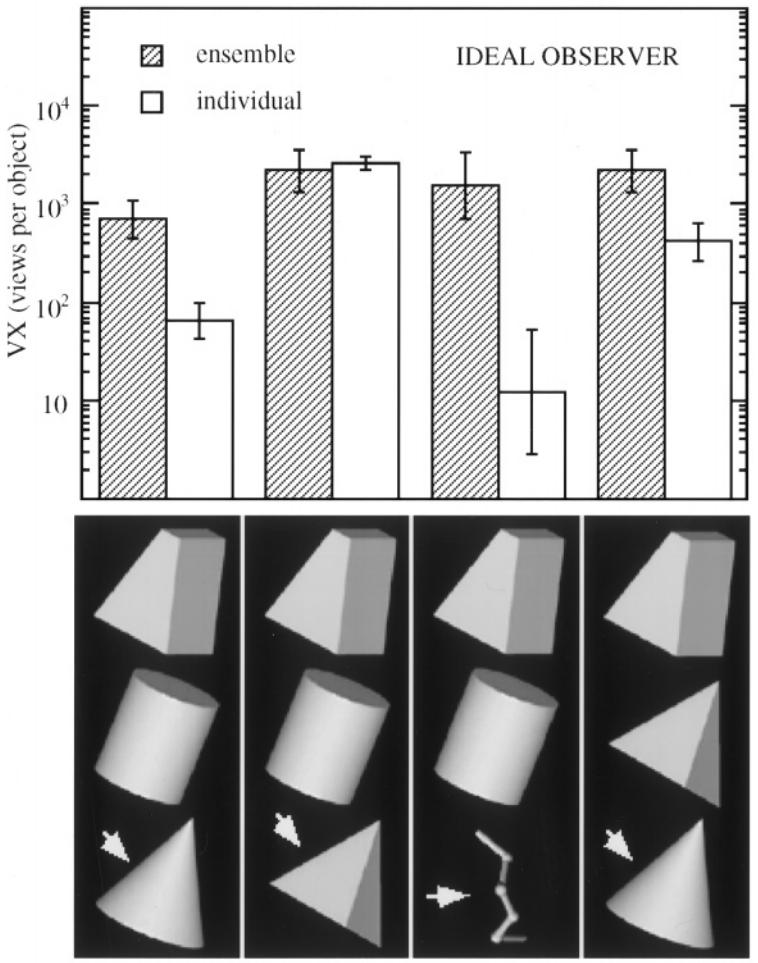

Fig. 9. Single-object VX measured for different objects in the same context (cone, pyramid or wire in the context of wedge and cylinder) and the same object in different contexts (cone in either wedge and cylinder, or wedge and pyramid). The task VX for each ensemble is plotted as shaded bars for comparison.

measurement we defined, which is luminance-noise based, could not be consistent with human data. However, when experiments are run at high contrast such that the visual system is not limited by low-level luminance noise, a VX measurement based on a structural-noise model could predict human performance.

A second limitation is on the number of objects that an be studied in a task, owing to practical computational constraints. More objects means more views to be considered in a simulation. To maintain estimation accuracy, the simulation time and memory space will have to grow in proportion to the square of the number of views involved. This is why all of our tasks involved sets of only three objects.

5.2. Viewpoint invariance and context

Can we characterize the intrinsic viewpoint invariance of an object in terms of how similar its views are to one another? Consider the notion of an aspect graph [28], which partitions the viewing sphere of an object into regions (aspects) bounded by 'visual events'. A visual event occurs when projective discontinuities (edges, corners, *etc*.) appear or disappear. Views within an aspect are considered 'perceptually similar'. An obvious use of the aspect-graph theory to assess the intrinsic viewpoint invariance of an object is to count the number of aspects of the object. The greater the number of aspects, the less intrinsic viewpoint invariance there is.

This type of approach is problematic in two ways. First, the definition of 'similarity' is based on certain assumptions about the visual system (*e.g*. metrical differences are less salient than projective discontinuities), which is not necessarily dictated by the task. As a result, it does not allow a clean separation of task properties from observer properties. Second, by only concerning within-object similarities, these single-object view-similarity measurements omit the context of a recognition task.

Strictly speaking, the intrinsic viewpoint invariance of a single object is undefined without a context. This

theoretical point becomes obvious when one tries to formulate an optimal classifier for object recognition. Minute differences between views can be crucial for discriminating objects in one context but irrelevant in another context. In the studies of viewpoint effects on human object recognition, the issue of context has often been ignored. The implicit assumption seems to be that context does not affect performance. This, however, is contrary to the general finding that scene context facilitates recognition speed and accuracy [29–31] and that perceptual similarity of objects is context sensitive [32–34].

5.3. VX and perceptual similarity

Psychological theories of object recognition often explain human performance in terms of perceptual similarity, a distance measurement in a 'psychological space' [35,33]. These theories do not deal directly with viewpoint effects, but by extension, they would predict increased viewpoint dependency when within-object similarity of views is low compared to between-object view similarity. Using parameterized animal shapes, Edelman [17] provided empirical support for this relationship between similarity and viewpoint effects in a discrimination task involving two classes of objects.

Although it is intuitive to say that viewpoint effects depend on the relative difference of within- and between-object similarities, to quantify this difference in terms of the two types of similarities turns out to be difficult. One reason is that the notions of within- and between-object similarities are difficult to define. Nosofsky [32], for example, has argued that similarity is context dependent. This means that the within-object similarity of one object is undefined unless other objects in a recognition task are known. It also means that within- and between-object similarities are inseparable. Another source of difficulty is the multi-dimensional nature of similarity. View A can be equally 'similar' to views B and C but along different dimensions (*e.g.* A is a trapezoid, B is a trapezoid with the same base and top as A but has a different height, and C is a trapezoid with a wider base and top than A but the same height). To quantify the 'relative difference' between within- and between-object similarities, one must take this multi-dimensional nature of similarity into account.

VX can be thought of as providing a single-number summary of the relative difference in within- *versus* between-object similarities, without having to directly measure these two types of similarities. A high VX, which we take to mean that a large number of views per object are needed to approximate the decision boundaries, suggests that views from different objects are not much more distinct than views from the same object. As currently defined, VX concerns similarities in the image space, which is the very first representation of the stimuli. If we formulate a VX measurement based on a plausible perceptual representation scheme of objects, then with respect to the assumed representation, the perceptual difference of within- and between-object similarities can be also assessed.

5.4. VX and human data

A high-VX task requires many image views to define the decision boundaries between objects. This means that each view is representative over a small area on the viewing sphere. Given this task property, if a human observer learns the objects from one view and tries to recognize them from a novel view, performance (speed and accuracy) is likely to deteriorate quickly as the difference in viewing angle between the learned and novel views increases, a hallmark of a viewpoint effect.

The VX measurements in Section 4.1 qualitatively correlate with human data on viewpoint effects in object recognition. We found low VX values for the simple geometric objects (single geons) and the mechanical parts (very distinct multiple-geon objects). This is consistent with the demonstrations that such objects yield viewpoint-invariant performance for human subjects [12]. Our wire and pea-like objects had high values of VX, consistent with the reported viewpoint-dependent performance in humans [16]. For faces, when an ecological constraint was imposed on orientation, we found a moderate VX, higher than that of the geometric objects but lower than that of the wires. Results in the human literature suggest that face recognition is viewpoint dependent [21–23], although to a lesser extent than that of the wire objects.

The correspondence between VX and human performance suggests that some important properties of human performance are due to the inherent viewpoint properties of the stimuli. Effects on human performance owing to perceptual representation are thus confounded by the VX of the stimuli. A definitive conclusion about human perceptual representation may have to await better control of the VX of stimuli.

5.5. The view-rate hypothesis

If correspondence between VX and human performance is a statement on how task properties dictate human performance, then a lack of correspondence between the two reveals the functioning of the human visual system. Biederman and Gerhardstein [12] studied charm bracelet objects similar to those used in Section 4.4. They found that human performance was viewpoint invariant, suggestive of a low VX. However, the VX of these charm bracelets is actually as high as that of the wire objects (>25000 views per object), which are known to produce viewpoint-dependent performance in humans. How do we account for this inconsistency between VX and human performance?

Suppose the speed of human visual processing is limited. Further suppose that this limit can be described in terms of a maximum equivalent number of views the system can process per unit time, where a view is defined in the input space of a task (*i.e.* an image). This hypothesis does not require that the visual system use a pure image-based processing scheme, but only that the speed at which intermediate representations are constructed, encoded and recognized can be characterized by a unit (views per second) defined in the image space. We call this hypothesis the view-rate hypothesis.

If the visual system were to maximize accuracy in all contrast and noise conditions, it would achieve recognition by implementing a Bayesian classifier that maximized the *a posterior* probability. The processing time of such an ideal-like observer would be monotonically related to the total-VX (VX × number of objects) of the task. This is because it would have to consider the equivalent of total-VX number of views before making a decision. However, there may be ecological pressures on the visual system to complete certain recognition tasks within a fixed amount of time. Given a view-rate bottleneck, the pragmatic visual system segments the problem into complex and less-complex components[7]. It will ignore the high-VX components when processing time is insufficient. Effectively, what are being recognized are not the original set of objects, but their lower-VX components. Compared with an ideal observer, the visual system gains speed but loses accuracy because of the added decision stage of segmentation and the ignoring of informative parts (albeit of high complexity).

The processing scenario described above suggests that the observed recognition time and viewpoint effect are dictated by the total-VX of the selected components (perceived total-VX), which is only indirectly related to the total-VX of the task (the original set of objects). This is apparently the case for the charm bracelets: the high-VX wire parts are ignored *a priori* and recognition depends only on a single geon-part. Because the recognition of isolated geons has a low VX, performance is viewpoint invariant.

Recently, Tarr *et al.* [36] studied charm bracelets with more than one charm. The charms in each bracelet were selected from a set of ten geons, such that no single geon by itself was diagnostic. Even though geons may still be the object parts used by the visual system in this case, the perceived total-VX is likely to be higher. This is because not only the identity of the geons, but also their spatial arrangement, must be identified. Tarr *et al.* found that human performance with these multi-charm bracelets is viewpoint dependent.

Two predictions can be made from the view-rate hypothesis. First, for a class of similar recognition tasks, for which there is good reason to believe that the same object components are being used by the visual system (*e.g.* face recognition under various shadow, lighting and viewpoint conditions [37]), recognition time and viewpoint dependence should increase monotonically with the total-VX of the task. This is because the perceived total-VX, which we hypothesized to be the determining factor for human performance, will be monotonic to the task total-VX if the same object components are being used across tasks.

The second prediction concerns the trade-off between the ability to generalize across views and statistical efficiency [38,39] of recognition. Recognition efficiency measures a human observer's ability to utilize information in a task to maximize accuracy [25]. It is the ratio of an ideal observer's threshold SNR to the human's threshold SNR. The prediction is that for two recognition tasks with similar total-VX (*e.g.* the wire-object task *vs.* the charm-bracelet task), if humans show more viewpoint invariance on one task, they will also be less efficient on that task. This is because viewpoint dependency is less for the task with a smaller perceived total-VX. However, smaller perceived total-VX means that fewer object components are being used by the visual system, resulting in a lower recognition efficiency. For the same reason, if the recognition time of two tasks with vastly different total-VX are about the same (*e.g.* geometric-object *vs.* charm-bracelet task), then recognition efficiency for the higher-VX task must be lower than the lower-VX task. Currently, both of these predictions await empirical confirmation.

6. Conclusion

VX is a metric for assessing the granularity of perceptual representation required to achieve a certain level of accuracy in an object-recognition task by an ideal observer. This measurement is independent of any visual processing and provides a way of gauging the amount of viewpoint invariance inherent in an object-recognition task. We described how the VX of a task can be reliably measured using Monte Carlo simulation. Our results illustrate how VX related to object shape, object components, performance criterion and context.

We found that the VX of recognition tasks can vary over a wide range, depending on what objects are to be recognized. This variation partially explains the inconsistency in human data regarding viewpoint effects on object recognition. To draw conclusions about perceptual representation, these data must be re-interpreted, taking VX into account.

[7] The term 'object component' is intended to be a more general concept than 'object part', which often refers to a structural element of an object. An object component can be any attribute about an object obtainable from an earlier processing stage. It can be color, spatial frequency component, object part or even structure of parts.

By comparing the VX's obtained using different sets of objects with the corresponding human data reported in the literature, we found suggestive evidence that the visual system may choose to represent informative object components or structures that are low in VX. We postulate that this is due to a limit in the equivalent number of views that the human visual system can process per unit time. This limit is called the view rate of the visual system. We made two predictions that can be used to test this hypothesis.

Acknowledgements

This work was supported by NIH Grant EY02857 to Gordon E. Legge. An earlier report appeared as a chapter in Bosco S. Tjan's doctoral thesis [40]. We thank Zili Liu for his helpful comments, Jeff Liter for suggestions on statistics and Bernhard Schölkopf for discussions on statistical learning theory.

Appendix A. Implementation of the ideal observer

In general, there is no analytical solution for obtaining threshold SNR for the ideal decision function of Eq. (1). Monte Carlo simulation was used to estimate this threshold. This was done by presenting a set of randomly chosen test targets (test views) to a computer implementation of the ideal observer over a fixed number of trials, each with a different noise pattern. In each trial, the computer made the optimal decision based on Eq. (1). The test views were randomly sampled from all object (task) views[8] (T_{ij}) allowed for the f-task, based on their prior probability distribution $p(T_{ij})$. In a trial, a stimulus was formed by adding a target to a freshly generated gaussian luminance noise pattern. To estimate the SNR required for a given accuracy criterion, the same block of trials was repeated at different values of SNR by changing the target contrast. A binary search procedure was used to locate the SNR that yields the threshold accuracy criterion.

All our tasks consisted of three objects, each with eight to 24576 views. The image size was 128×128 pixels. A brute-force implementation of this simulation would require full evaluation of Eq. (1) at every trial. For our recognition tasks, it would be too slow (even on a fast workstation, the simulations could take months to complete). Significant speed-up is, however, possible at the expense of memory space. This was the route we took. The most time-consuming part of evaluating Eq. (1) is computing the square of the Euclidean distance, $\|R - T_{ij}\|^2$. We used the following approach to speed up this computation. Let T^0_{ij} denote a 'standard' template of object i view j, which has its background luminance set to zero. Let N be a noise field of zero mean and unit variance, and B be a uniform field of background luminance[9]. The stimulus R and the template T_{ij} used in a trial are constructed as:

$$R = aT^0_{\mathrm{mn}} + B + \sigma N$$
$$T_{ij} = aT^0_{ij} + B \tag{A1}$$

The constant a determines the signal contrast and σ determines the noise power. They jointly determine the SNR of the stimulus R [$\mathrm{SNR} = (a^2/\sigma^2)(T^0_{\mathrm{mn}})^2$]. To avoid complicated notations, let X denote a standard view used to construct the stimulus (*i.e.* $X = T^0_{\mathrm{mn}}$) and T be a standard view for forming the template T_{ij} (*i.e.* $T = T^0_{ij}$). The square of the Euclidean distance calculation in Eq. (1) can be expressed as follows:

$$\begin{aligned}\|R - T_{ij}\|^2 &= R^2 - 2RT_{ij} + T^2_{ij} \\ &= (aX + \sigma N)^2 - 2(aX + \sigma N)aT + (aT)^2\end{aligned} \tag{A2}$$

By rearranging teems and deleting those that do not depend on i or j, we obtain the following expression, which is monotonic in the square of the Euclidean distance between the stimulus and a template:

$$\begin{aligned}\|R - T_{ij}\|^2 &\underset{i,j}{\approx} -2(aX + \sigma N)aT + (aT)^2 \\ &\underset{i,j}{\approx} -2a^2XT + a\sigma NT + a^2TT\end{aligned} \tag{A3}$$

Notice that while X, T and N are image arrays, each of their dot-products, XT, NT and TT, is only a scalar number. Because they do not depend on the stimulus SNR, these dot-products can be pre-computed, saved in files and reused for every iteration in the binary search for the threshold SNR. Furthermore, changing the test targets (X's) involves recomputing only the XT term; changing the noise sample (N) requires recomputing only NT.

To obtain X's and T's, we need to render the objects from different viewpoints. Because of the large number of views needed, we have to render them on-demand from their 3-D models. The program uses

[8] Here we assume that the position of the center of object rotation is somehow normalized. If not, the additional spatial uncertainty can be handled by adding more views to account for all possible spatial locations confined by the image area. See the discussion section of Tjan et al. [25] for a formulation and simulation of an ideal observer with spatial uncertainty.

[9] The background luminance is not required for calculating SNR and does not affect an ideal observer's performance. It is here to avoid having to display negative luminance in the corresponding human experiment.

a caching scheme to maximize memory utilization and to reduce the number of times a view must be regenerated. Additional savings were achieved by first intersecting two views to eliminate zero-luminance background from the dot-product (multiply-and-add) calculation.

The intermediate files of the dot-products are huge. Each task described in this paper consisted of three objects, 24576 views each. There were 96 test views and 600 noise patterns. As a result, there were $96 \times 3 \times 24576 = 7.1$ million XT dot-products, $600 \times 3 \times 24576 = 44.2$ million NT's and $3 \times 24576 = 73728$ TT's. With the dot-products represented by eight-byte double-precision numbers, the file sizes were 54 million, 338 million and 576000 bytes, respectively. On an SGI Power Indigo 2 graphics workstation, it took about 2–3 days to generate these files. Using these files, the subsequent binary search for threshold SNR took < 20 min for each block of 600 trials.

Our ideal observer simulation strategy can be summarized as follows: given the task views of an f-task (which specifies T's), a set of test views (X's) and a set of noise patterns (N's), we created three intermediate files containing the dot-products XT, NT and TT. When completed, these files were read into (virtual) memory. The current SNR for a given block of trials was then set according to a binary search procedure. For each trial, the program determined the identity of the target object in a noisy image by evaluating Eq. (1), substituting Eq. (A3) for the square of the Euclidean distance.

References

[1] Tarr MJ, Bülthoff HH. Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein, 1993. J Exp Psychol: Hum Percep Perform 1995;21(6):1494–505.

[2] Biederman I, Gerhardstein PC. Viewpoint-dependent mechanisms in visual object recognition: reply to Tarr and Bülthoff, 1995. J Exp Psychol: Hum Percep Perform 1995;21(6):1506–14.

[3] Marr D, Nishihara HK. Representation and recognition of the spatial organization of three-dimensional shapes. Proc R Soc Lond B 1978;200:269–94.

[4] Biederman I. Recognition by components: a theory of human image understanding. Psychol Rev 1987;94(2):115–47.

[5] Ullman S. Aligning pictorial descriptions: an approach to object recognition. Cognition 1989;32:193–254.

[6] Sirovich L, Kirby M. Low-dimensional procedure for the characterization of human faces. J Opt Soc Am A 1987;4(3):519–24.

[7] Poggio T, Edelman S. A network that learns to recognise 3-D objects. Nature 1990;343(6255):263–6.

[8] Buhmann J, Lange J, von der Malsburg C. Distortion invariant object recognition by matching hierarchically labeled graphs. In: IJCNN: International Joint Conference on Neural Networks, Washington, DC, vol. 1. New York, NY: IEEE TAB Neural Network Committee, 1989:155–159.

[9] Ullman S, Basri R. Recognition by linear combination of models. IEEE Trans Pattern Anal Machine Intelligence 1991;13(10):992–1006.

[10] Biederman I, Cooper EE. Evidence for complete translational and reflectional invariance in visual object priming. Perception 1991;20(5):585–93.

[11] Cooper EE, Biderman I, Hummel JE. Metric invariance in object recognition: a review and further evidence. Can J Psychol 1992;46(2):191–214.

[12] Biederman I, Gerhardstein PC. Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. J Exp Psychol: Hum Percep Perform 1993;19(6):1162–82.

[13] Tarr MJ, Pinker S. Mental rotation and orientation dependence in shape recognition. Cogn Psychol 1989;21(2):233–82.

[14] Rock I, DiVita J. A case of viewer-centered object perception. Cogn Psychol 1987;19(2):280–93.

[15] Bülthoff HH, Edelman S. Psychophysical support for a 2-D view interpolation theory of object recognition. Proc Natl Acad Sci USA 1992;1989(1):60–4.

[16] Edelman S, Bülthoff HH. Orientation dependence in the recognition of familiar and novel views of 3-D objects. Vis Res 1992;32(12):2385–400.

[17] Edelman S. Class similarity and viewpoint invariance in the recognition of 3-D objects. Biol Cybernet 1995;72:207–20.

[18] Gauthier I, Tarr MJ. Becoming a 'Greeble' expert: exploring mechanisms for face recognition. Vis Res 1997;37(12):1673–82.

[19] Humphrey GK, Shakeela CK. Recognizing novel views of 3-D objects. Can J Psychol 1992;46(2):170–90.

[20] Liu Z. Viewpoint dependency in object representation and recognition. Spat Vis 1996;9(4):491–521.

[21] Krouse FL. Effects of pose, pose change and delay on face recognition performance. J Appl Psychol 1981;66:651–4.

[22] Logie RH, Baddeley AD, Woodhead MM. Face recognition, pose and ecological validity. Appl Cogn Psychol 1987;1:53–69.

[23] Troje NF, Bülthoff HH. Face recognition under varying poses: the role of texture and shape. Vis Res 1996;36(12):1761–71.

[24] Duda RO, Hart PE. Pattern Classification and Scene Analysis. New York: Wiley, 1973.

[25] Tjan BS, Braje WL, Legge GE, Kersten D. Human efficiency for recognizing 3-D objects in luminance noise. Vis Res 1995;35(21):3053–69.

[26] Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer, 1995.

[27] Press WH, Teukolsky SA, Vetterling WT, Flannery BPI. Numerical Recipes in C, 2nd edn. Cambridge, UK: Cambndge University Press, 1992.

[28] Koenderink JJ, van Doom AJ. The internal representation of solid shape with respect to vision. Biol Cybernet 1979;32:211–6.

[29] Curry R, Hung GK, Wilder J, Julez B. Context effect of common objects on visual processing. Optom Vis Sci 1995;72(7):452–60.

[30] De Graef P, De Troy A, D'Ydewalle G. Local and global contextual constraints on the identification of objects. Can J Psychol 1992;46(3):480–508.

[31] Henderson JM. Object identification in context: the visual processing of natural scenes. Can J Psychol 1992;46(2):319–41.

[32] Nosofsky RM. Choice, similarity, and context theory of classification. J Exp Psychol: Learn Mem Cogn 1984;10(12):104–14.

[33] Ashby FG, Perrin NC. Toward a unified theory of similarity and recognition. Psychol Rev 1988;95(1):124–50.

[34] Cutzu F, Edelman S. Faithful representation of similarities among 3-D shapes in human vision. Proc Natl Acad Sci USA 1996;93(21):12046–50.

[35] Nosofsky RM. Attention, similarity, and the identification-categorization relationship. J Exp Psychol: Gen 1986;115(1):39–61.
[36] Tarr MJ, Bülthoff HH, Zabinski M, Blanz V. To what extent do unique parts influence recognition across changes in viewpoint? Psychol Sci 1997;8(4):282–9.
[37] Tjan BS, Kersten D, Braje WB. Inherent illumination invariance in face recognition. Invest Ophthalmol Vis Sci 1997;38(4):S1002.
[38] Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd, 1925.
[39] Tanner WP, Birdsall TG. Definitions of d' and η as psychophysical measures. J Opt Soc Am 1958;30(10):922–8.
[40] Tjan BS. Ideal observer analysis of object recognition. Doctoral Thesis, Computer Science Department, University of Minnesota, 1996.