THE ROYAL SOCIETY

Received 3 May 2002
Accepted 21 June 2002
Published online 29 August 2002

Recognizing novel three-dimensional objects by summing signals from parts and views

David H. Foster* **and Stuart J. Gilson†**

Visual and Computational Neuroscience Group, Department of Optometry and Neuroscience, University of Manchester Institute of Science and Technology, Manchester M60 1QD, UK

Visually recognizing objects at different orientations and distances has been assumed to depend either on extracting from the retinal image a viewpoint-invariant, typically three-dimensional (3D) structure, such as object parts, or on mentally transforming two-dimensional (2D) views. To test how these processes might interact with each other, an experiment was performed in which observers discriminated images of novel, computer-generated, 3D objects, differing by rotations in 3D space and in the number of parts (in principle, a viewpoint-invariant, 'non-accidental' property) or in the curvature, length or angle of join of their parts (in principle, each a viewpoint-dependent, metric property), such that the discriminatory cue varied along a common physical scale. Although differences in the number of parts were more readily discriminated than differences in metric properties, they showed almost exactly the same orientation dependence. Overall, visual performance proved remarkably lawful: for both long (2 s) and short (100 ms) display durations, it could be summarized by a simple, compact equation with one term representing generalized viewpoint-invariant parts-based processing of 3D object structure, including metric structure, and another term representing structure-invariant processing of 2D views. Object discriminability was determined by summing signals from these two independent processes.

Keywords: three-dimensional object discrimination; visual shape recognition; structural descriptions; geometrical invariants; templates; depth rotations

1. INTRODUCTION

As we move about the world and view objects from different directions and distances, the images falling on the retina undergo corresponding changes in shape and position. How, despite rarely experiencing the same image twice, do we recognize the same objects and distinguish between different ones? Since the seminal work of Pitts & McCulloch (1947), explanatory theories have concentrated either on visually extracting properties of the image that do not depend on viewpoint (an invariants approach) or on visually compensating for changes in the image as viewpoint changes (a transformational approach). The work done in the first visual process can be regarded as a trade-off against that done in the second (Shepard 1975).

Invariants-based approaches have concentrated on local rather than global geometric properties and have assumed, typically, that objects are represented visually as structural descriptions, consisting of primitive parts, for example, cylinders and spheres, and the spatial relations between those parts, such as 'connected to' and 'left of' (Sutherland 1968; Barlow *et al.* 1972; Marr & Nishihara 1978; Hoffman & Richards 1984; Biederman 1987; Logothetis & Sheinberg 1996; Wu & Levine 1997). These descriptions are inferred from a range of essentially viewpoint-invariant and non-invariant properties of the image; for example, connectivity and collinearity are strictly preserved over all viewpoints, except for occlusions and other peculiarities of view; parallelism is preserved only where perspective effects are negligible; and planar curvature, a metric property, is largely preserved only where rotations and translations in depth are small (e.g. Lowe 1985; Binford & Levitt 1993). In general, 'non-accidental' image properties are those that remain stable over a range of viewpoints, and, insofar as they are not an accident of view, can provide reliable information about three-dimensional (3D) structure (Binford 1981; Biederman 1987; Dickinson *et al.* 1997).

By contrast, transformation-based approaches have assumed, typically, that objects are represented visually in a view-specific way, as two-dimensional (2D) templates or 'views', specifying properties such as the metric coordinates of their constituent points or local features (Ullman 1989; Bülthoff & Edelman 1992) to which internal restoring or normalizing transformations, such as rotations, translations and dilatations are applied, but at a cost proportional to the angular difference in view (Shepard & Metzler 1971; Foster & Mason 1979; Tarr *et al.* 1998; cf. Willems & Wagemans 2001). Comprehensive viewpoint-invariant recognition can be achieved only if additional views are made available, from which other instances can be recovered by interpolation or extrapolation (Bülthoff & Edelman 1992) or other linear combination (Ullman & Basri 1991).

Both approaches have been elaborated considerably (e.g. Cutzu & Edelman 1998; Tarr & Bülthoff 1998; Biederman & Bar 1999), in particular, the specification of the conditions under which parts-based theories might produce viewpoint invariance (Biederman & Bar 2000; Hayward & Tarr 2000) and how views might be defined (Tarr & Kriegman 2001). Both approaches have also partially converged, with the addition of metric information to parts-based descriptions (e.g. Hummel & Stankiewicz 1998) and the addition of structural information to view-

* Author for correspondence (d.h.foster@umist.ac.uk).
† Present address: University Laboratory of Physiology, University of Oxford, Parks Road, Oxford OX1 3PT, UK.

based descriptions (e.g. Tarr & Bülthoff 1998; Edelman & Intrator 2000). Because of these developments, it may sometimes be difficult to decide whether a given experimental task involves parts-based or view-based processing (Wagemans *et al.* 1996; Vanrie *et al.* 2001).

The approach taken here to the question of the processes underlying object recognition was neutral: rather than attempting to isolate experimentally one kind of processing or the other, the aim was to provide a framework within which both might be identified, to elucidate their characteristics and to determine how each contributes to observed performance.

The experimental task required the discrimination of novel 3D objects differing by randomly chosen rotations in 3D space. Objects were generated by computer in such a way that individual aspects of their structures, including a strict invariant, namely the number of parts, and three metric properties, namely curvature, length and angle of join of parts, could be systematically manipulated along a common, physically defined, dimensionless scale (cf. Foster 1980; Foster & Ferraro 1989; Hummel & Stankiewicz 1996). Observers were not trained on particular views of objects nor exposed to multiple views other than the two images for comparison, which were presented simultaneously to reduce the confounding effects of memory. Although cast as a discrimination task under 3D rotations, it can also be interpreted as a matching task (e.g. Lawson & Humphreys 1996; Hayward & Williams 2000), in which recognition across time is exchanged for recognition across space. To test whether viewpoint invariance was immediate or whether shifts in eye fixations were necessary to make these judgements (Just & Carpenter 1976), both long- and short-duration image displays were used.

It was found that object parts were treated distinctly: detecting differences in their number was much easier than detecting differences in their metric properties; yet the two kinds of discrimination had almost exactly the same viewpoint dependence, declining as the angular orientation difference increased to 45°, but then remaining constant and well above chance. Critically, the effects of manipulating structure and orientation difference did not interact. Performance could be accurately summarized by a simple, compact equation, which, it is suggested, represents the activity of generalized parts-based and view-based processes, the signals from which are summed to determine object discriminability.

2. MATERIAL AND METHODS

(a) *Stimuli*

Images of objects were presented on the screen of a high-resolution computer graphics display system. Several factors determined the design of the objects. To avoid the confounding effects of prior knowledge, familiarity and semantic content, all of which may vary from observer to observer, objects were generated afresh in each trial by a solid-rendering process that produced realistic images of 3D structures based on a random walk. To minimize the effects of self-occlusion with medium-to-large rotations in depth without imposing bilateral symmetry, objects were sparsely structured and were formed by concatenating, at variable angles, cylinders with axes of variable curvature and length, as illustrated in figure 1. Details of the stimuli and of the

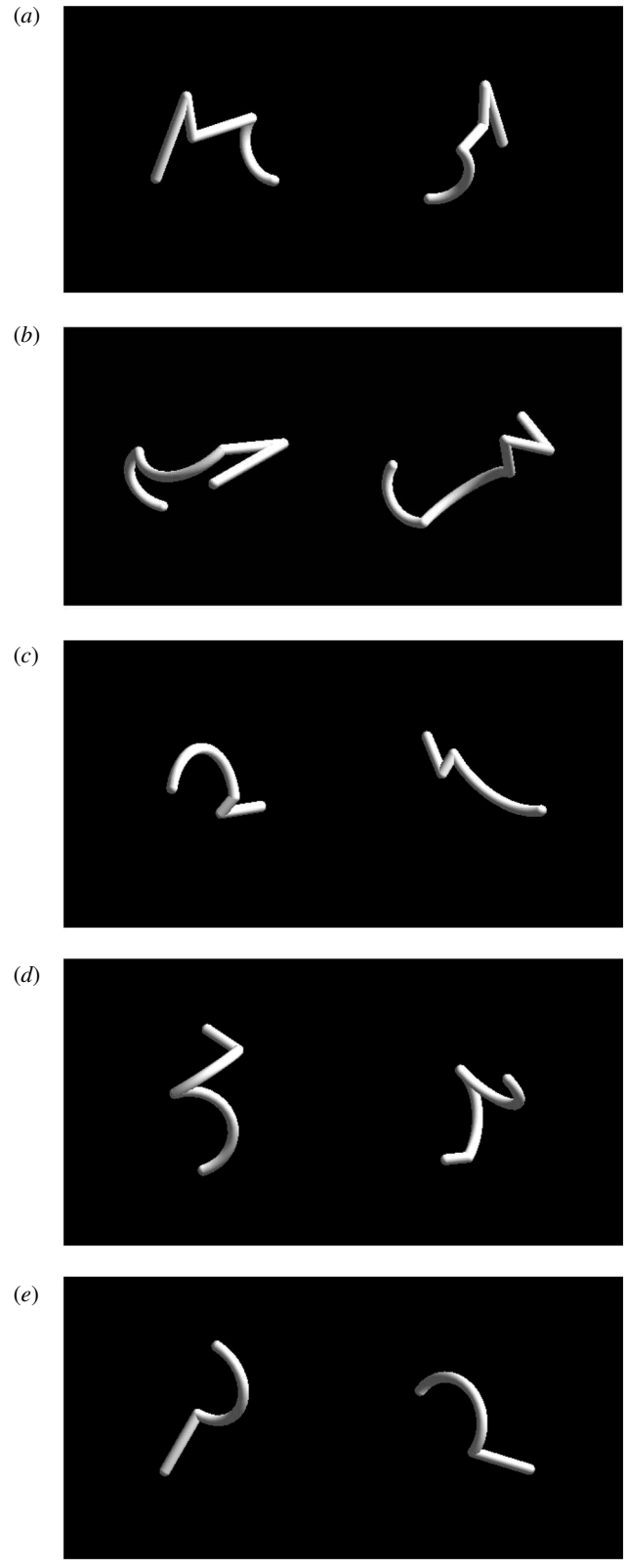

Figure 1. Example images of (*a*) 'same' and (*b*)–(*e*) 'different' pairs of randomly generated stimulus objects drawn from 28 800 pairs used in the experiment. The objects in 'same' pairs differed only in 3D orientation. The objects in 'different' pairs differed both in orientation and in (*b*) the number of parts; (*c*) the curvature of one of the parts; (*d*) the end-to-end length of one of the parts; or (*e*) the angle between two parts. Differences in object orientation and structure varied randomly from trial to trial. Each image subtended a visual angle no greater than 5.5°.

graphics display system are given in Appendix A. These objects differed from the straight-axis, constant-segment-length 'paperclip' cylinder figures used in some object-recognition studies, which have occasionally been criticized for producing similar structural descriptions (e.g. Biederman & Gerhardstein 1993, 1995; Tarr & Bülthoff 1995). By construction, the cylinders were assumed to define the 'parts' of the objects. Evidence that parts-based processing occurred is presented later, but, in principle, the problem of recovering 3D parts-based descriptions of complex objects from single intensity images seems to be tractable, for example, by exploiting the projective properties of curved-axis cylinders (e.g. Zerroug & Nevatia 1996, 1999).

The structure of the objects was varied by varying the number of parts in each object and their (continuous) metric properties, that is, the curvature and length of the axis of each part and the relative orientation of one part in relation to the next. From previous work on psychophysically efficient descriptors of planar curvature (Foster *et al.* 1993), the deviation from linearity ('sag') rather than Euclidean curvature was used to quantify the curvature of each axis. As already indicated, axis curvature in an image is not itself a strict invariant, except when reduced to a binary value signifying 'straight' or 'curved' (Lowe 1985; Biederman 1987); nor are length and relative orientation. But the number of parts in an image—discounting occlusions—is such an invariant, and therefore a strong non-accidental property; indeed, theoretically, it is the most basic of all invariants (Bourbaki 1968, ch. 4; Foster 1975).

In each trial, observers were presented with images of the 'same' or 'different' pairs of objects, each object subtending no more than 2.5° visual angle and separated by at least 0.5° visual angle. 'Same' pairs of objects were produced identically (figure 1*a*). 'Different' pairs were produced identically except with respect to one of the four properties of their parts; that is, they differed in the number of cylinders (figure 1*b*) or in the curvature of one of the cylinders (figure 1*c*) or in the length of one of the cylinders (figure 1*d*) or in the angle between two cylinders (figure 1*e*). For both 'same' and 'different' pairs, one of the objects in the pair was given a rotation in three dimensions by an angle θ chosen randomly from the range 0°, 15°, …, 345°. Each rotation θ was about the centroid of the object, and whether it was in the fronto-parallel plane or not was chosen randomly (i.e. each with $p = 0.5$). If it was not, the axis of revolution in 3D space was chosen randomly and without constraint. (A subsequent analysis of variance showed no significant effect of whether the rotation was in the fronto-parallel plane.) Whether a 'same' or 'different' pair appeared was chosen randomly (i.e. each with $p = 0.5$). Display duration was either 2 s or 100 ms, the latter too short for planned changes in fixation. The screen was viewed binocularly at 2 m in a moderately darkened room. Viewing conditions were photopic.

(b) *Procedure*

The observer initiated each trial and, after the image had been presented, responded as to whether the pair of objects was the same or different, irrespective of viewpoint, by pressing an appropriate key on a keyboard connected to the computer. If the pair was a 'different' pair, the distinguishing property and its value were chosen randomly, so that neither object properties nor orientation differences were blocked (cf. Biederman & Bar 1999; Hayward & Williams 2000), and therefore observers could not anticipate the nature of the cue, if any. Trials were performed in sessions lasting no more than 1 h, with short breaks between each block of 120 trials. In each session, only long-duration or only short-duration displays were used. Each observer was informed before the experiment and reminded before each session about the nature of the stimuli, the ways in which the objects could differ and the equal probability of 'same' and 'different' trials. Observers were encouraged to respond as quickly as was consistent with accuracy and were not given feedback, except as a total percentage correct at the end of each block. Over a period of several months, each observer performed 1800 trials with long-duration and with short-duration displays. Details of the experimental design are given in Appendix A. There was little effect of long-term learning (mean performance levels in the second half of trials were *ca.* 7% higher than in the first half, for both long- and short-duration stimuli; $F_{1,15} \leqslant 4.4$; $p > 0.05$).

There were 16 observers, each with Snellen acuity better than 6/6, aged 18–35 years, and, except for one (co-author S.J.G.), they were unaware of the purpose of the experiment and were paid for their participation.

(c) *Analysis*

The natural scales of all four properties—number, curvature, length and angle of join of the cylinders—were incommensurate. Cue values defined by differences along these four scales were therefore re-expressed along a common dimensionless scale obtained by dividing by a normalizing factor (Atkinson & Donev 1992). In principle, these normalizing factors could have been either global, for example, the maximum of the range of values of the property defined over all trials, or local, for example, the mean of the values of the property within the trial, as in a Michelson contrast. For the purpose of data presentation, the maximum was used; that is, for any pair of values c_j and c_k of some property ranging over values c_i, the value Δc of the cue was defined as $\Delta c = (c_j - c_k)/\max\{c_i\}$, where $c_j > c_k$. As demonstrated later, using other measures, including Michelson contrast did not alter the pattern of results. It is emphasized that the reason for using a physically defined scale was not to ensure that equal steps along the scale were necessarily equally salient for different object properties, which can be difficult to establish and may obscure important uniformities in response; rather, the aim was simply to make possible the comparison of differences in salience of equal physical steps along the scale for different object properties, much as one might compare differences in sensitivity to changes in luminance and colour for equal steps along a physical scale quantified by Michelson contrast (for more general discussion, see Falmagne (1985)). The advantages of this procedure become clear later.

Discrimination performance was measured by the discrimination index d' from signal-detection theory (Green & Swets 1966) rather than by response time, thus emphasizing encoding-level processes rather than decision-level processes (Rouder 2000). This index has several advantages in the present context. It linearizes and combines responses to 'same' and 'different' trials, thereby minimizing the effects of observer bias; it eliminates the compression that occurs near the limits of a percentage-correct scale; and it is additive (Durlach & Braida 1969). In brief, let HR be the hit rate, that is, the proportion of 'different' responses to different-object pairs; let FAR be the false-alarm rate, that is, the proportion of 'different' responses to same-object pairs; and let z be the inverse of the cumulative unit normal distribution; then $d' = z(\mathrm{HR}) - z(\mathrm{FAR})$. A zero value of d' corresponds to chance performance and d' increases monotonically with the detectability of the target (in a two-alternative forced-choice task, a value of d' of 1.0 corresponds to 76%

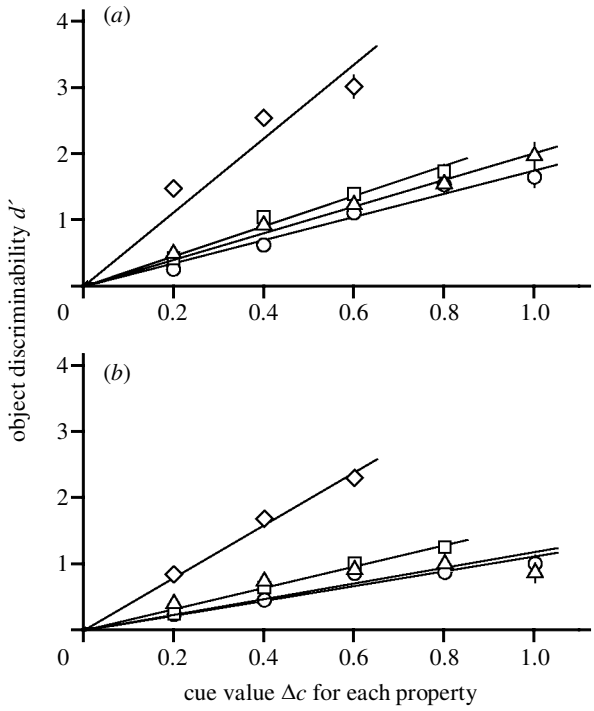

Figure 2. Discriminability of objects differing in number, curvature, length or angle of join of parts. Discrimination index d' (symbols) calculated over 16 observers is plotted against normalized cue value Δc for each of the four properties (diamonds, number; circles, curvature; squares, length; triangles, angle). Vertical bars show ± 1 s.e., where these are sufficiently large. The straight lines through each set of points are linear regressions constrained to pass through the origin. Data for (*a*) 2 s and (*b*) 100 ms display durations were each based on 28 800 trials.

correct). Individual performances by observers were similar and hit and false alarm scores were pooled over observers.

3. CUES FOR SHAPE DISCRIMINATION

How do differences in object structure affect object discriminability? Figure 2 shows discrimination index d', pooled over all observers and orientation differences, plotted against cue value Δc for each of the four object properties: number, curvature, length and angle of join of parts. Data in figure 2*a* are for 2 s displays and in figure 2*b* for 100 ms displays. Mean response time (RT) was 0.7 s for 2 s displays and 1.1 s for 100 ms displays; there was no trade-off between d' and RT: the linear regression of RT on d' was -0.068 ± 0.012 s for 2 s displays and -0.006 ± 0.017 s for 100 ms displays (values are given ± 1 s.e.).

For both display durations, discrimination based on number was markedly different from discriminations based on curvature, length and angle of join, the latter giving closely similar values. The data were well approximated by a linear dependence of d' on Δc, although the largest value of d' may have been limited by a ceiling effect. The values of the gradients are listed in table 1. The ratio of the gradient for number to mean gradient for curvature, length and angle of join was 2.8 and 3.0 for long- and short-duration displays, respectively. The gradients scaled almost uniformly with change in display duration, with a mean ratio of *ca*. 0.7.

Table 1. Gradients of discrimination index d' against normalized cue value Δc for each of four object properties: number of parts, curvature, length and angle of join of parts.

	number	curvature	length	angle
2 s display	5.57	1.75	2.27	2.01
100 ms display	3.97	1.12	1.61	1.19
ratio	0.71	0.64	0.71	0.59

Table 2. Gradients of discrimination index d' against Michelson cue value Δc for each of four object properties: number of parts, curvature, length and angle of join of parts.

	number	curvature	length	angle
2 s display	8.29	1.49	2.63	1.86
100 ms display	5.70	1.05	1.80	1.35
ratio	0.69	0.71	0.69	0.73

This separation in linear dependencies for number and for curvature, length and angle of join of parts is unlikely to be an artefact of the method of normalizing each range of values, as the following control procedures showed.

(i) The normalization of cue values was changed from global to local. Thus, with the cue value Δc defined by Michelson contrast, $\Delta c = (c_j - c_k)/(c_j + c_k)$, discrimination index d' increased less smoothly with Δc, but, as table 2 shows, the ratio of gradient for number to mean gradient for curvature, length and angle of join was still high: 4.2 and 4.1 for long- and short-duration displays, respectively.

(ii) The experiment was repeated with non-uniform changes in the ranges for each property. Thus, the range for number was reduced to 2, 3; for curvature, to 0.0, 3.5, 7.0 mm; for length, to 7.0, 14.0, 21.0 mm; and for angle of join, to −60°, −30°, ..., 60°. A different group of observers took part and only short-duration displays were used. With the now smaller normalizing factors $\max\{c_l\}$ in the first definition of cue value, $\Delta c = (c_j - c_k)/\max\{c_l\}$, for each property, the ratio of gradient for number to mean gradient for curvature, length and angle of join was 3.3, almost the same as with the original ranges.

(iii) The gradients for number, curvature, length and angle of join were calculated when one member of the pair contained, in turn, just 2, 3 and 4 parts. The corresponding ratios of the gradient for number to mean gradient for curvature, length and angle of join were 2.5, 3.0 and 3.5 for long-duration displays, and 2.4, 2.7 and 3.5 for short-duration displays. The separation in linear dependencies increased rather than decreased as the number of parts in the objects increased, and therefore could not be attributed to discrimination dominated by objects with the fewest parts (cf. Biederman & Bar 1999).

4. VOLUME-BASED SHAPE DISCRIMINATION

It might be argued that the difference between discriminations based on number of parts and on metric properties

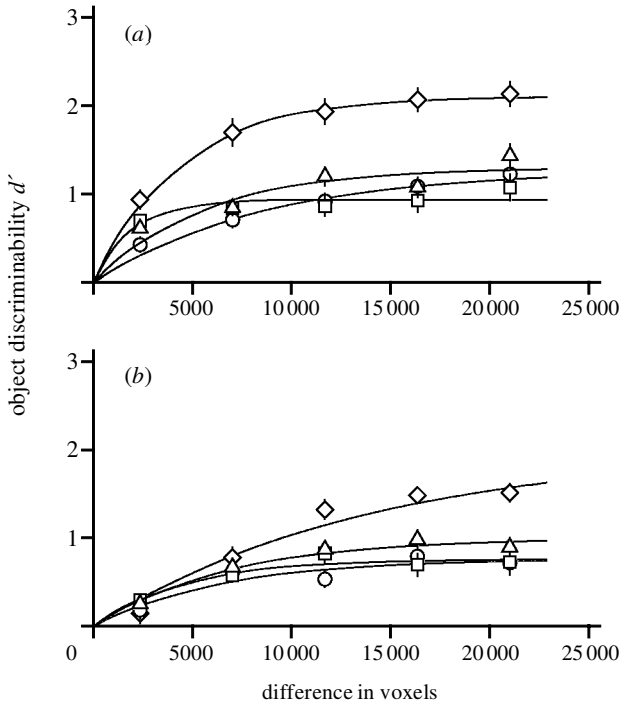

Figure 3. Discriminability of objects differing in 3D overlap. Discrimination index d' (symbols) is plotted against difference in voxels between pairs of objects after translating and rotating the objects in three dimensions to produce maximum overlap. Data are for (*a*) 2 s and (*b*) 100 ms displays with discriminations based on differences in number (diamonds), curvature (circles), length (squares) or angle of join (triangles) of parts. Vertical bars show ± 1 s.e., where these are sufficiently large. The smooth curves through each set of points are best-fitting saturating exponential functions. Voxel calculations were based on 9600 out of 28 800 pairs of objects used in the experiment, with values pooled within bins of size of 4660 voxels.

is due simply to volumetric differences in the objects from which the images were derived; in other words, equal steps along the normalized stimulus ranges produced larger differences in object volume when the number of parts was varied than when metric properties were varied. To test this hypothesis, voxel representations of the objects in each pair were generated from those used in the experiment. They were then translated and rotated in three dimensions to produce maximum overlap. The number of voxels in the symmetric difference (XOR) of the result was recorded, along with the observer's 'same' or 'different' response to the original image pair. This computationally intensive operation was limited to 9600 out of 28 800 pairs of images seen by observers. Figure 3 shows observed values of discrimination index d' plotted against voxel difference for number, curvature, length and angle of join of parts, binned over intervals of 4660 voxels determined by the unit of volume used to define the objects. Data in figure 3*a* are for 2 s displays and in figure 3*b* for 100 ms displays. The effects of object property remained: identical differences in voxels produced different levels of discrimination depending on whether they were due to differences in numbers of parts or in metric properties.

An alternative, less plausible notion is that the difference between the discriminations is due to differences in the 2D distributions of light associated with these manipulations. To test this hypothesis, the images in each pair used in the experiment were binarized, translated to produce maximum overlap, and the number of pixels in the symmetric difference of the result recorded, along with the observer's 'same' or 'different' response to the original image pair. As with voxel-based calculations, the effects of property remained, despite identical pixel differences.

5. EFFECTS OF VIEW

How does the difference in 3D orientation between the two objects affect their discriminability? Figure 4 shows discrimination index d', pooled over all observers and cue values, plotted against angular orientation difference θ for number, curvature, length and angle of join of parts. Data in figure 4*a* are for 2 s displays and in figure 4*b* for 100 ms displays. Performance was best with identical object orientations $\theta = 0°$ and it declined as θ increased up to *ca.* 45° (cf. Bülthoff & Edelman 1992), but it did not decline to chance levels. In fact, over most of the angular range, 60–300°, performance remained high and almost constant for both display durations, with levels of d' ranging from 0.6 to 1.7, depending on display duration and on which property provided the discriminatory cue.

This constant performance over 60–300° cannot be attributed to viewpoint-invariant processing of the simplest, two-part images (cf. Biederman & Bar 1999) and viewpoint-dependent processing of the rest; for discrimination performance varied with θ in the same way for two-part images alone. It also cannot be attributed to averaging different patterns of viewpoint-dependent discrimination over successive object pairs. To see this, suppose that a given pair of objects produces a high d' score at some non-zero angular orientation difference, $\theta = \theta_0$ say, as well as at $\theta = 0°$ (recall that as each object pair was presented in the experiment only once, information about the same pair at different angles was unavailable). Although this score is unlikely to be as high as at 0°, it ought to decline with θ in the same way. Thus, suppose that d' varies with θ about θ_0 according to some function, say g, in the same way as it does about 0°, but scaled down uniformly by a factor a, say; that is, $g(\theta) = ag((\theta - \theta_0)/a)$, for all θ in some neighbourhood of θ_0. Then a simple calculation shows that averaging viewpoint-dependent discrimination over object pairs with randomly varying θ_0 yields a mean that varies from 0.4 to less than 0.1 of that observed over 60–300° as a varies from 0.75 to 0.25. Other functions g produce other means, but to account for the observed level of d' over 60–300° most object pairs need to produce viewpoint-invariant discrimination over most orientation differences.

The orientation dependence of discrimination of number of parts over the interval 0–45° (and 315–360°) was not unexpected (Tarr *et al.* 1998; and, for analogous discrimination of planar patterns, Foster (1978)). It might be argued, however, that although objects were designed to minimize occlusion, accidental views, for example, where a foreshortened axis appeared straight at one orientation but curved at another, were sufficiently common that orientation costs dominated performance. Although such accidental effects may have influenced some discriminations, it seems unlikely that they were decisive.

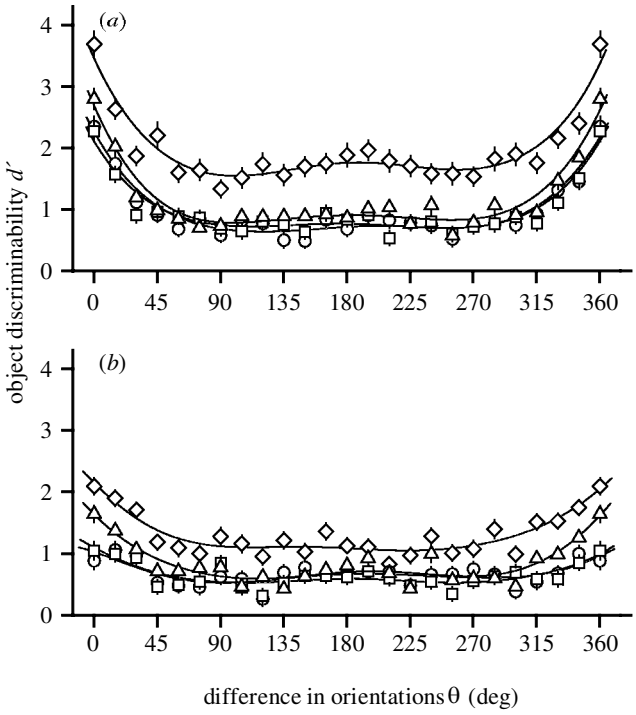

Figure 4. Discriminability of objects differing in orientation. Discrimination index d' (symbols) calculated over 16 observers is plotted against angular orientation difference θ between 3D-rotated objects differing in number (diamonds), curvature (circles), length (squares) or angle of join (triangles) of parts. The points plotted at 360° duplicate those at 0°. Vertical bars show ± 1 s.e., where these are sufficiently large. The smooth curves through each set of points are locally weighted quadratic regressions. Data for (*a*) 2 s and (*b*) 100 ms display durations were each based on 28 800 trials.

First, for discrimination based on curvature, the orientation dependencies were not significantly different for objects with five parts and those with two parts ($F_{3,20} = 0.40$, $p > 0.5$), although the probability of at least one axis being foreshortened would have been greater for objects with five parts. Second, as figure 4 shows, apart from a shift in d' level, the effect of difference in angular orientation was almost exactly independent of the property providing the cue, except perhaps near $\theta = 0°$ for discrimination based on angle of join. A formal statistical analysis of variance confirmed this general inference: for both long- and short-duration stimuli, there was no significant interaction between effects of cue type and angular orientation difference ($F_{69,759} \leqslant 0.97$, $p \geqslant 0.5$); that is, the effects of these two factors were additive (cf. Tarr *et al.* 1997).

A broader analysis of variance showed significant interactions between cue type and cue size for long- and short-duration stimuli ($F_{6,48} = 4.4$, $p = 0.001$; $F_{6,18} = 4.0$, $p = 0.01$, respectively) and between the effects of angular orientation difference and cue size for long-duration stimuli ($F_{14,112} = 2.28$, $p < 0.01$), but not for short-duration stimuli. As with the effects of cue size (figure 2), the effects of angular orientation difference scaled almost uniformly with reduction in display duration, the ratio being *ca.* 0.7 (cf. Lawson & Humphreys 1996).

6. COMBINING PARTS AND VIEWS

The linear dependence of discrimination index on cue value shown in figure 2 (§ 3) and the additivity of the effects of angular orientation difference and cue type shown in figure 4 (§ 5) suggest the possibility of a simple, compact description of visual performance. For a given display duration of, say, 2 s, suppose that the gradient relating the linear dependence of d' on cue value Δc for cue type i is k_i (given by the entries in table 1, with $i = 1$ for number of parts, 2 for curvature, etc.), that is, $d' = k_i \Delta c$, and suppose that the nonlinear dependence of d' on angular orientation difference θ is $f(\theta)$ (derived from the mean over i of the graphs in figure 4), that is, $d' = f(\theta)\Delta c$. Then these two d'-values can be summed to produce the observed level of discrimination performance

$$d' = [k_i + f(\theta)]\Delta c. \qquad (6.1)$$

The first term in brackets accounts for the viewpoint-invariant dependence of discrimination performance on object structure; the second term for the structure-invariant dependence of discrimination performance on viewpoint. Notice that, although there is no interaction between the effects of i and θ, there are interactions between the effects of i and Δc and between the effects of θ and Δc. In practice, k_i can be assumed to have just two values, one for number of parts and the other, about one-third the size, for metric properties. The effect of decreasing display duration from 2 s to 100 ms is accommodated by uniformly multiplying k_i and $f(\theta)$ by a constant, *ca.* 0.7 from table 1.

In view of the three-factorial design of the experiment and the potential for confounds, this equation gives a remarkably accurate account of the data: for long-duration stimuli, the proportion of variance accounted for was 91% (root mean square error = 0.26, residual d.f. = 84); for short-duration stimuli, the proportion of variance accounted for was 88% (root mean square error = 0.21, residual d.f. = 84).

An interpretation of the equation in terms of visual processes is suggested in the next section.

7. DISCUSSION

For the novel 3D objects considered here, visual discrimination depended on both object structure and 3D orientation, but not in the way that might have been anticipated. Discrimination based on number of parts was different from discrimination based on curvature, length and angle of join of parts; and it was not merely better for a given cue size, it increased more rapidly with cue size by a factor of *ca.* 3. This advantage was not an artefact of the choice of scale used to measure the non-accidental and metric properties of parts, or of differences in the distributions of light over the retina or in the volumes of the underlying 3D structures. Despite the importance of parts, viewpoint had a marked effect on performance: as the difference in angular orientation between two objects increased, discrimination performance decreased for angles up to *ca.* 45° independent of which property provided the discriminatory cue. Beyond these limits, performance remained almost constant and well above chance level, thereby defining a large range, 60–300°, of true viewpoint-invariant discrimination.

Importantly, the absence of interaction between cue type and orientation difference suggests that the actual viewpoint invariance of some theoretically non-accidental properties need not be greater than that of metric properties (Tarr *et al.* 1998), provided that an appropriate linear measure of performance is used and orientation differences are not blocked (cf. Biederman & Bar 1999). This result should be distinguished from that reported in Hayward & Williams (2000) showing the absence of interaction between task difficulty and angular orientation difference, over angles of 0–60°.

The fact that highly reliable discrimination could be achieved with novel images of 100 ms duration is consistent with processing being immediate and not requiring point-by-point shifts in eye fixation or attention (Just & Carpenter 1976; Tarr *et al.* 1997; Biederman & Bar 1999). As the dependence of discrimination performance on stimulus parameters had the same form for long- and short-duration displays, it seems that observers were able to extract object information without the application of high-level attentive effort, or extensive training. With regard to methodological issues, it also indicates that no particular difficulty was experienced in representing the kinds of stimuli used here, and, in particular, parsing them into parts.

As discrimination based on each of the three metric properties of curvature, length and angle of join could be described by a common equation, it seems that the device of expressing performance with respect to the kinds of physically defined dimensionless scale used here helps eliminate some of the unessential aspects of object structure, thereby allowing more general kinds of visual representation to be probed (Ishai *et al.* 1999). The pattern of performance summarized by equation (6.1) does not of course imply that, in other circumstances, other factors may not be decisive; for example, viewpoint invariance can be achieved trivially by adding non-geometric cues such as colour (Tarr & Bülthoff 1995; Hayward & Williams 2000). For more everyday objects under 3D rotations, prior knowledge, familiarity and semantic content also influence observed performance.

As already intimated, the dependence of discrimination performance on object structure and orientation difference, summarized by equation (6.1), is consistent with the simultaneous operation of two independent processes (Logothetis & Sheinberg 1996). One process is independent of viewpoint but dependent on structure, and the other is dependent on viewpoint but independent of structure. These two processes are not, however, completely identical with the parts-based and view-based processes outlined in § 1. Thus, the first, viewpoint-invariant, structure-dependent process is parts-based, that is sensitive to parts, but in a generalized way so that all the properties of parts, both non-accidental and metric, are processed in a viewpoint-invariant fashion over all orientations, and not just over the visible interval of 60–300°. This is computationally feasible as viewpoint-invariant metric properties could be estimated directly from invariants and quasi-invariants extracted from the image (e.g. Zerroug & Nevatia 1999). The second, viewpoint-dependent, structure-invariant process is view-based, but also in a generalized way so that the similarity or difference between objects, whether in metric or non-metric properties, is determined by the point-wise similarity or difference in their 2D images. The results of these two processes then combine additively to determine observed performance. Notice that the combination is unlikely to have been competitive, as in a race model. Taking the maximum of the corresponding discriminatory signals rather than their sum implies an interaction between the effects of cue type and orientation difference that was not observed.

The notion of two independent processes acting additively to achieve 3D object discrimination under rotations in depth may have a partial analogy with a combination of discrete and continuous processes that was suggested for 2D pattern discrimination under translations and rotations in the plane (Foster & Kahn 1985). This discrete process was assumed to operate on parts or local features and their spatial relations and the continuous process on metric properties with an efficiency that declined with the extent of the compensatory translation or rotation required. Both operations appeared necessary to achieve the observed levels of discrimination performance (Foster 1991). In both two and three dimensions, therefore, the visual system may use the same, limited repertoire of shape-processing strategies.

APPENDIX A

(a) *Stimulus construction*

Each object consisted of several concatenated straight or curved cylinders; that is, the axis of each cylinder was straight or an arc of a variable circle in the plane, and the cross-section of the cylinder was a constant circle, locally orthogonal to its axis. The diameter of the cross-section was 3.5 mm and the variable end-to-end length of each cylinder (i.e. the chord length of the axis) was no greater than 35.0 mm. Spheres of diameter 3.5 mm were inserted at the ends of each cylinder to provide a smooth join. Objects were generated in real-time as a random walk in 3D space, starting at the origin, and constrained to avoid self-intersection and excursion beyond the bounding volume. The objects, which were white and presented in an almost black field, were rendered with a Phong illumination model and Gouraud interpolation between vertices. The direction of the illumination appeared to be from above the observer's right shoulder. The images were presented under orthographic projection. Adding perspective produced the same pattern of observer performance.

From pilot experiments, the number of cylinders in each object was chosen randomly from the range 2–5, which provided a useful distribution of discrimination levels and covered the range for immediate perception of number and possibly a little beyond ('subitization': Taves 1941; Atkinson *et al.* 1976; Luck & Vogel 1997). The end-to-end length of each cylinder was chosen randomly from the range 7.0, 10.5, …, 35.0 mm. The angle between adjoining cylinders at the point of join was chosen randomly from the range −150°, −120°, …, 150°. The remaining two angles (the second and third Euler angles) completed the specification of the orientation of one cylinder relative to the other, and were also chosen randomly and without constraint. Whether each cylinder was straight or curved was chosen randomly (i.e. each with $p = 0.5$). The curvature of each cylinder (i.e. of its axis) was parameterized by its sag (Foster *et al.* 1993), the value of which was chosen

randomly from the range 3.5, 7.0, …, 17.5 mm, but constrained so that it did not exceed half the cylinder length.

(b) *Experimental design*

The number of unique 'different' spatially ordered pairs of values for each property was 12, 110, 20 and 110 for number, curvature, length and angle of join, respectively. So that cues based on each property appeared equally often, pairs of values were packaged into sequences of 1320 trials; thus, the 12 unique number pairings were duplicated 110 times, the 110 curvature pairings 12 times, and so on. When added to an equivalent number of 'same' trials, this procedure produced a sequence of 10 560 trials. Distributed over a minimum of six observers (actually 16 participated), each had to perform 1760 trials, which over blocks of 120 trials rounded up to 1800 trials, the ordering of which was fully randomized. That this distribution corresponds to 37.5 trials for each orientation difference had no special significance.

As the values of each property were uniformly sampled (i.e. each number of parts occurred equally often, each curvature value equally often, and so on), small differences occurred more frequently than large differences. The fact that the discriminations based on curvature, length and angle of join produced closely similar linear dependencies of discrimination index on cue value suggests that these differences in frequencies were unimportant.

(c) *Graphics display system*

Stimuli were displayed with a spatial resolution of 1280×1024 pixels and 24-bit colour on the screen of a 21-inch RGB monitor (model CM2198MSG; Hitachi, Yokohama). This was controlled by a specialist graphics workstation with a high-speed graphics subsystem for real-time accurate simulation of 3D images (Onyx 2 Reality Station; SGI, Mountain View, CA). Images were anti-aliased. As a control on the fidelity of the stimuli, they were photographed at their display durations and the photographic images measured and compared with their specifications. The screen refresh rate was 60 Hz, and measured 90–10% decay and 10–90% rise times of the phosphors were less than 1 ms.

The authors thank S. S. Baker for assistance in running experimental sessions, H. H. Bülthoff for helpful discussion, and C. Christou, R. Lawson and E. Pauwels for comments on the manuscript. This work was supported by the Engineering and Physical Sciences Research Council.

REFERENCES

Atkinson, A. C. & Donev, A. N. 1992 *Optimum experimental designs*. Oxford: Clarendon.

Atkinson, J., Campbell, F. W. & Francis, M. R. 1976 The magic number 4 ± 0: a new look at visual numerosity judgements. *Perception* **5**, 327–334.

Barlow, H. B., Narasimhan, R. & Rosenfeld, A. 1972 Visual pattern analysis in machines and animals. *Science* **177**, 567–574.

Biederman, I. 1987 Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* **94**, 115–147.

Biederman, I. & Bar, M. 1999 One-shot viewpoint invariance in matching novel objects. *Vis. Res.* **39**, 2885–2899.

Biederman, I. & Bar, M. 2000 Differing views on views: response to Hayward and Tarr (2000). *Vis. Res.* **40**, 3901–3905.

Biederman, I. & Gerhardstein, P. C. 1993 Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *J. Exp. Psychol. Hum. Percept. Perform.* **19**, 1162–1182.

Biederman, I. & Gerhardstein, P. C. 1995 Viewpoint-dependent mechanisms in visual object recognition: reply to Tarr and Bülthoff (1995). *J. Exp. Psychol. Hum. Percept. Perform.* **21**, 1506–1514.

Binford, T. O. 1981 Inferring surfaces from images. *Artif. Intell.* **17**, 205–244.

Binford, T. O. & Levitt, T. S. 1993 Quasi-invariants: theory and exploitation. In *Proc. 22nd DARPA image understanding workshop, Washington, DC, 18–21 April 1993* (ed. O. Firschein), pp. 819–829. San Francisco, CA: Morgan Kaufmann.

Bourbaki, N. 1968 *Theory of sets*. Reading, MA: Addison–Wesley.

Bülthoff, H. H. & Edelman, S. 1992 Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl Acad. Sci. USA* **89**, 60–64.

Cutzu, F. & Edelman, S. 1998 Representation of object similarity in human vision: psychophysics and a computational model. *Vis. Res.* **38**, 2229–2257.

Dickinson, S. J., Bergevin, R., Biederman, I., Eklundh, J.-O., Munck-Fairwood, R., Jain, A. K. & Pentland, A. 1997 Panel report: the potential of geons for generic 3-d object recognition. *Image Vis. Comput.* **15**, 277–292.

Durlach, N. I. & Braida, L. D. 1969 Intensity perception. I. Preliminary theory of intensity resolution. *J. Acoust. Soc. Am.* **46**, 372–383.

Edelman, S. & Intrator, N. 2000 (Coarse coding of shape fragments) plus (retinotopy) approximate to representation of structure. *Spatial Vis.* **13**, 255–264.

Falmagne, J.-C. 1985 *Elements of psychophysical theory*. Oxford: Clarendon.

Foster, D. H. 1975 An approach to the analysis of the underlying structure of visual space using a generalized notion of visual pattern recognition. *Biol. Cybern.* **17**, 77–79.

Foster, D. H. 1978 Visual comparison of random-dot patterns: evidence concerning a fixed visual association between features and feature-relations. *Q. J. Exp. Psychol.* **30**, 637–654.

Foster, D. H. 1980 A spatial perturbation technique for the investigation of discrete internal representations of visual patterns. *Biol. Cybern.* **38**, 159–169.

Foster, D. H. 1991 Operating on spatial relations. In *Pattern recognition by man and machine* (ed. R. J. Watt), pp. 50–68. Houndmills, UK: Macmillan.

Foster, D. H. & Ferraro, M. 1989 Visual gap and offset discrimination and its relation to categorical identification in brief line-element displays. *J. Exp. Psychol. Hum. Percept. Perform.* **15**, 771–784.

Foster, D. H. & Kahn, J. I. 1985 Internal representations and operations in the visual comparison of transformed patterns: effects of pattern point-inversion, positional symmetry, and separation. *Biol. Cybern.* **51**, 305–312.

Foster, D. H. & Mason, R. J. 1979 Transformation and relational-structure schemes for visual pattern recognition. *Biol. Cybern.* **32**, 85–93.

Foster, D. H., Simmons, D. R. & Cook, M. J. 1993 The cue for contour-curvature discrimination. *Vis. Res.* **33**, 329–341.

Green, D. M. & Swets, J. A. 1966 *Signal detection theory and psychophysics*. New York: Wiley.

Hayward, W. G. & Tarr, M. J. 2000 Differing views on views: comments on Biederman and Bar (1999). *Vis. Res.* **40**, 3895–3899.

Hayward, W. G. & Williams, P. 2000 Viewpoint dependence and object discriminability. *Psychol. Sci.* **11**, 7–12.

Hoffman, D. D. & Richards, W. A. 1984 Parts of recognition. *Cognition* **18**, 65–96.

Hummel, J. E. & Stankiewicz, B. J. 1996 Categorical relations in shape perception. *Spatial Vis.* **10**, 201–236.

Hummel, J. E. & Stankiewicz, B. J. 1998 Two roles for attention in shape perception: a structural description model of visual scrutiny. *Visual Cogn.* **5**, 49–79.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L. & Haxby, J. V. 1999 Distributed representation of objects in the human ventral visual pathway. *Proc. Natl Acad. Sci. USA* **96**, 9379–9384.

Just, M. A. & Carpenter, P. A. 1976 Eye fixations and cognitive processes. *Cogn. Psychol.* **8**, 441–480.

Lawson, R. & Humphreys, G. W. 1996 View specificity in object processing: evidence from picture matching. *J. Exp. Psychol. Hum. Percept. Perform.* **22**, 395–416.

Logothetis, N. K. & Sheinberg, D. L. 1996 Visual object recognition. *A. Rev. Neurosci.* **19**, 577–621.

Lowe, D. G. 1985 *Perceptual organization and visual recognition*. Boston, MA: Kluwer Academic.

Luck, S. J. & Vogel, E. K. 1997 The capacity of visual working memory for features and conjunctions. *Nature* **390**, 279–281.

Marr, D. & Nishihara, H. K. 1978 Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond.* B **200**, 269–294.

Pitts, W. & McCulloch, W. S. 1947 How we know universals. The perception of auditory and visual forms. *Bull. Math. Biophys.* **9**, 127–147.

Rouder, J. N. 2000 Assessing the roles of change discrimination and luminance integration: evidence for a hybrid race model of perceptual decision making in luminance discrimination. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 359–378.

Shepard, R. N. 1975 Form, formation, and transformation of internal representations. In *Information processing and cognition: the Loyola symposium* (ed. R. L. Solso), pp. 87–122. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shepard, R. N. & Metzler, J. 1971 Mental rotation of three-dimensional objects. *Science* **171**, 701–703.

Sutherland, N. S. 1968 Outlines of a theory of visual pattern recognition in animals and man. *Proc. R. Soc. Lond.* B **171**, 297–317.

Tarr, M. J. & Bülthoff, H. H. 1995 Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993) *J. Exp. Psychol. Hum. Percept. Perform.* **21**, 1494–1505.

Tarr, M. J. & Bülthoff, H. H. 1998 Image-based object recognition in man, monkey and machine. *Cognition* **67**, 1–20.

Tarr, M. J. & Kriegman, D. J. 2001 What defines a view? *Vis. Res.* **41**, 1981–2004.

Tarr, M. J., Bülthoff, H. H., Zabinski, M. & Blanz, V. 1997 To what extent do unique parts influence recognition across changes in viewpoint? *Psychol. Sci.* **8**, 282–289.

Tarr, M. J., Williams, P., Hayward, W. G. & Gauthier, I. 1998 Three-dimensional object recognition is viewpoint dependent. *Nat. Neurosci.* **1**, 275–277.

Taves, E. H. 1941 Two mechanisms for the perception of visual numerousness. *Arch. Psychol.* **37**, 1–47.

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

Ullman, S. & Basri, R. 1991 Recognition by linear combinations of models. *IEEE Trans. Patt. Anal. Mach. Intell.* **13**, 992–1006.

Vanrie, J., Willems, B. & Wagemans, J. 2001 Multiple routes to object matching from different viewpoints: mental rotation versus invariant features. *Perception* **30**, 1047–1056.

Wagemans, J., Van Gool, L. & Lamote, C. 1996 The visual system's measurement of invariants need not itself be invariant. *Psychol. Sci.* **7**, 232–236.

Willems, B. & Wagemans, J. 2001 Matching multicomponent objects from different viewpoints: mental rotation as normalization? *J. Exp. Psychol. Hum. Percept. Perform.* **27**, 1090–1115.

Wu, K. & Levine, M. D. 1997 3-d shape approximation using parametric geons. *Image Vis. Comput.* **15**, 143–158.

Zerroug, M. & Nevatia, R. 1996 Three-dimensional descriptions based on the analysis of the invariant and quasi-invariant properties of some curved-axis generalized cylinders. *IEEE Trans. Patt. Anal. Mach. Intell.* **18**, 237–253.

Zerroug, M. & Nevatia, R. 1999 Part-based 3d descriptions of complex objects from a single image. *IEEE Trans. Patt. Anal. Mach. Intell.* **21**, 835–848.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.