PERGAMON

Vision Research 41 (2001) 1981–2004

Vision Research

www.elsevier.com/locate/visres

What defines a view?☆

Michael J. Tarr [a,*], David J. Kriegman [b]

[a] *Department of Cognitive and Linguistic Sciences, Brown University, Box 1978, Providence, RI 02912, USA*
[b] *Department of Computer Science, Beckman Institute, University of Illinois at Urbana-Champaign, Champaign, IL 61801, USA*

Received 9 September 2000; received in revised form 10 October 2000

Abstract

At a given instant we see only visible surfaces, not an object's complete 3D appearance. Thus, objects may be represented as discrete 'views' showing only those features visible from a limited range of viewpoints. We address how to define a view using Koenderink's (Koenderink & Van Doorn, Biol. Cybernet. 32 (1979) 211.) geometric method for enumerating complete sets of stable views as *aspect graphs*. Using objects with known aspect graphs, five experiments examined whether the perception of orientation is sensitive to the qualitative features that define aspect graphs. Highest sensitivity to viewpoint changes was observed at locations where the theory predicts qualitative transitions, although some transitions did not affect performance. Hypotheses about why humans ignore some transitions offer insights into mechanisms for object representation.

Keywords: Object representation; Viewpoint effects; Shape perception; Characteristic views; Aspect graphs

1. Introduction

> If I hold my head to the left and look down at the handle grips and front wheel and map carrier and gas tank I get one pattern of sense data. If I move my head to the right I get another slightly different pattern of sense data. The two views are different. The angles of the planes and curves of the metal are different. The sunlight strikes them differently. If there's no logical basis for substance then there's no logical basis for concluding that what's produced these two views is the same motorcycle.
>
> (Pirsig, 1974).

2. Human object representation

A common assumption is that visual object recognition is accomplished by comparing what we see to visual/spatial mental representations of objects. Inherent in this model of recognition are three issues: first, what is the format of such representations; second, how are such representations acquired; and, third, what are the mechanisms and processes used to match input images to such representations. For the most part, it has been the first issue, the format of object representations, that is central to theories of visual cognition. In addition to subserving recognition, such representations are held to be common to mental simulations of physical phenomena, visual problem solving, and spatial reasoning (Shepard & Cooper, 1982). Thus, elucidating the specific properties of the representations underlying object recognition may increase our understanding of many facets of visual cognition.

While it is possible to draw many distinctions between different representational formats for objects, one of the most commonly held is between *viewpoint-dependent* and *viewpoint-independent* representations. Consistent with this distinction are two approaches to object representation: multiple-views and structural-descriptions. The former is based on viewpoint-specific features and images (Poggio & Edelman, 1990; Bülthoff & Edelman, 1992; Tarr, 1995), while the latter is based on configurations of three-dimensional (3D) parts (Marr, 1982; Biederman, 1987; Hummel &

☆ This work was originally presented at the *Workshop on Visual Perception: Computation and Psychophysics, Cape Cod, MA, January, 1993.*

* Corresponding author. Tel.: +1-401-863-1148; fax: +1-401-863-2255.

E-mail address: michael_tarr@brown.edu (M.J. Tarr).

0042-6989/01/$ - see front matter
PII: S0042-6989(01)00024-4

Biederman, 1992). Certain elements from both approaches are likely to play some role in human object recognition in that both viewpoint-dependent and viewpoint-independent patterns of recognition performance have been obtained (Jolicoeur, 1985; Bülthoff & Edelman, 1992; Biederman & Gerhardstein, 1993; Tarr, 1995). Notwithstanding these more global differences, there are certain commonalties between the two approaches. Both approaches propose that self-occlusions within an object lead to multiple representations. As features become visible or occluded with changes in viewpoint, each 'new' configuration is instantiated as a distinct representation; referred to as a 'view' or 'aspect' in some models and simply an 'alternate' representation in others. The essential question underlying this phenomenon and the central issue addressed by this article is: *What properties of the object define a view?* At the most general level, one answer agreed upon by both approaches is that new representations across changes in viewpoint are defined *qualitatively*, that is, by a change in particular features, usually geometric, derived from the image (Biederman & Gerhardstein, 1995 and Tarr & Bülthoff, 1995 debate). It is important to note that qualitative change is a relative concept meaningful only for a given class of features, a change that may be considered qualitative for one class may not be qualitative for another. Thus, one difference between the multiple-views and structural-description approaches is what constitutes an appropriate feature set for organizing the representation (Biederman and Gerhardstein (1995), Tarr & Bülthoff debate different potential feature sets). What we introduce here is one computational model of how shape features may vary qualitatively across changes in viewpoint. We then use psychophysical methods to test whether human observers are sensitive to the qualitative transitions specified by this particular approach.

2.1. Empirical studies of object representation

To provide a better understanding of why the selection of a particular qualitative feature set is important for theories of recognition, we first present a brief review of some of the most relevant experimental results. A fundamental question is how recognition may be achieved across variation in the two-dimensional (2D) image arising from changes in 3D viewpoint. Although both multiple-views and structural-description theories propose that gross occlusions are compensated for by the encoding of more than one view, they differ in their predictions for how perceivers generalize from familiar viewpoints of objects to unfamiliar viewpoints *within* views. Consistent with the dichotomy drawn between theories, multiple-views theories suggest that views are specific to familiar viewpoints and that recognition performance, in terms of both response time and accuracy, will decrease with increasing distance between familiar and unfamiliar viewpoints (Tarr & Pinker, 1989; Tarr, 1995). In contrast, many structural-description theories argue that a particular configuration of features will be invariant over changes in viewpoint and that recognition performance will remain *constant* with increasing distance between familiar and unfamiliar viewpoints (Biederman & Gerhardstein, 1993, 1995). Thus, these two approaches make different predictions regarding the *stability* of the features that define a given view and regarding how the features encoded within a given view are matched to input shapes.

Several studies provide evidence that, at least in some cases, human observers use multiple viewpoint-specific representations for object recognition. For example, both Tarr and Pinker (Tarr & Pinker, 1989; Tarr, 1995) and Bülthoff and Edelman (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992) familiarized participants with novel objects in a small set of viewpoints and then tested generalization to new, unfamiliar viewpoints. They found that the recognition of objects at new viewpoints was progressively worse as a function of the distance from the nearest familiar viewpoint. Supporting the idea that objects are learned as multiple view-based representations is an extensive body of work sharing the common empirical finding of viewpoint dependence in recognition (Rock, 1973; Bartram, 1974; Rock, 1974; Bartram, 1976; Palmer, Rosch, & Chase, 1981; Jolicoeur, 1985; Rock & Di Vita, 1987; Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Humphrey & Khan, 1992; Hayward & Tarr, 1997). For example, Bartram found that the time to name familiar objects decreased with practice more rapidly when the same viewpoint was repeatedly presented as compared to when new viewpoints were presented. In a subsequent study, Bartram found that the time to judge whether sequentially presented line drawings of objects were the same was faster for same-viewpoint images as compared to different-viewpoint images of the same object (see also Lawson, Humphreys, & Watson, 1994). For photographs, Bartram found a similar pattern for less familiar objects, but near viewpoint invariance for very familiar objects. This latter result may be interpreted in terms of multiple views. Familiar objects are more likely to have been seen in many viewpoints as compared to less familiar objects (Tarr & Pinker, 1989). Thus, comparisons across viewpoint between unfamiliar objects will require additional processing since the displayed viewpoints are more likely to be novel.

Recent findings from a range of disciplines also provide evidence for view-based representations. From a neuroscientific perspective, Perrett, Rolls, and Caan (1982), Perrett et al. (1989), Perrett et al. (1991) and Perrett, Oram, and Ashbridge (1998) have found cells in the monkey cortex that are sensitive to specific viewpoints of familiar objects such as faces. Likewise, Logothetis and colleagues (Logothetis & Pauls, 1995; Logothetis, Pauls, & Poggio, 1995) trained monkeys to recognize novel objects from specific viewpoints. In addition to replicating the pattern of viewpoint-dependent recognition behavior found in humans with similar objects (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992), they found evidence for arrays of view-tuned neurons in monkey cortex that corresponded to the trained viewpoints.

Plaut and Farah (1990) cite both neurophysiological (Perrett et al., 1985) and neuropsychological (Ratcliff & Newcombe, 1982) evidence in support qualitatively-defined view-based representations. They note that while cell responses are invariant with respect to image plane transformations, cells are sensitive to depth plane transformations – a pattern consistent with the hypothesis that view-based representations are structured according to qualitative changes in viewpoint. This proposal is also supported by studies of *agnosics*, i.e., recognition-impaired brain lesion patients, that sometimes exhibit selective impairment in recognizing different views of objects (Warrington & Taylor, 1973; Layman & Greene, 1988) – in particular, exhibiting difficulties with unconventional or less familiar views. From a computational perspective, Seibert and Waxman (1990) and Seibert and Waxman (1992) have developed a neural network system that uses multiple 2D views to represent and recognize 3D objects. This model utilizes a learning architecture that is based on biologically motivated models of neural interaction. Several other researchers have also attempted to implement multiple-views recognition models within neurally plausible networks (Poggio & Edelman, 1990; Weinshall, Edelman, & Bülthoff, 1990; Edelman & Weinshall, 1991; Hummel & Stankiewicz, 1996b) or other, more traditional, computational frameworks (Freeman & Chakravarty, 1980; Ullman, 1989; Ullman & Basri, 1991).

While the above studies examined the effect of viewpoint *in general*, Biederman and Gerhardstein (1993) and Hayward and Tarr (1997) have investigated a more specific question – how changes in the visible image structure of objects influence viewpoint dependency. In both studies participants decided whether two sequentially-presented objects were the same regardless of any rotation in depth. For multi-part objects, Biederman and Gerhardstein found that when the same set of parts remained visible across two viewpoints, generalization was much better as compared to when the set of visible parts changed between viewpoints. Biederman and Gerhardstein interpreted this result as evidence that objects are represented as structural-descriptions composed of viewpoint-invariant parts: when particular parts become visible or occluded a new structural-description must be activated. Hayward and Tarr, however, were able to obtain almost exactly the same pattern of results using single-volume objects. Their interpretation was that it is configurations of qualitative features, not parts that define different views of objects. Under either interpretation, both results suggest that different object representations are generated depending on the image structure that arises at a given viewpoint, a result consistent with current versions of both the multiple-views and structural-description approaches. Thus, despite differences in the specific features used to represent objects ranging from simple local image patches (Riesenhuber & Poggio, 1999) to qualitatively defined 3D parts (Biederman, 1987) almost all models of object recognition assume multiple view-based representations. What remains an open question is the principle by which views are organized. It is this problem we turn to next.

2.2. What is a view?

Despite the large number of studies that have examined viewpoint-dependent recognition and view-based representations, there has been little effort to elucidate precisely what constitutes a 'view.' Indeed, given the fact that humans are active observers that must acquire object representations over time,[1] one of the most important aspects of any view-based approach is the mechanism by which an observer determines whether two occurrences of an object are similar. This process necessarily includes several different pieces of information: the pose of the object; whether the object is familiar; and, whether the object has been observed at that viewpoint.[2] Moreover, using such information, the visual system must establish both the identity of the object and whether the current view should be retained as a new view (either as a distinct view or as a 'basis image' in the computation of a model of object shape) of an existing object representation, as an entirely new

[1] In contrast, typical computer-based recognition systems rely on three-dimensional object models that are specified a priori by the designer.

[2] Mechanisms for computing such information prior to recognition commonly employ a subset of the input image, for instance, a small number of orientation-free local features (Ullman, 1989). More generally, all recognition schemes must include *indexing* procedures in order to identify the most likely match between input and all known object models (Clemens & Jacobs, 1991). Such procedures may, as part of their estimate of fit, supply information about pose or familiarity.

object representation, or not at all. Indeed, studies of how observers instantiate views within object representations indicate that three factors play a role.

1. Familiarity with an object in a given viewpoint prompts the instantiation of a view (Tarr & Pinker, 1989; Bülthoff & Edelman, 1992; Tarr, 1995).
2. Familiarity with some exemplars from a class of visually-similar objects in a given viewpoint prompts the instantiation of a class-general view (Jolicoeur & Milliken, 1989; Lando & Edelman, 1995; Moses, Ullman, & Edelman, 1996; Tarr & Gauthier, 1998). Observers are able to use information about the particular objects seen at specific viewpoints to make inferences about the appearance of new members of the class seen only at a canonical viewpoint. These 'virtual views' (Poggio & Vetter, 1992; Beymer & Poggio, 1996) may arise through visual similarities between objects within a class (Gauthier & Tarr, 1997b), as well as symmetries within a specific object.
3. It has been found that the way in which the geometry of a given object changes with viewpoint may prompt the instantiation of new views (Vetter, Poggio, & Bülthoff, 1994; Tarr & Gauthier, 1998), in particular, with greater changes in image structure leading to an increased likelihood of a new aspect or 'characteristic' view[3] (Freeman & Chakravarty, 1980). For example, Edelman and Bülthoff (1992) found better recognition performance for some viewpoints of objects despite the fact that all of the novel test objects were seen equally often from each test viewpoint (a result consistent with the 'canonical views' phenomenon first reported by Palmer et al. (1981)). Moreover, these preferred views persisted even with extensive practice and the addition of depth cues (which might have facilitated 3D viewpoint-invariant representations).

It is this final factor, the way in which image structure changes with viewpoint, that is of the greatest relevance in defining what constitutes a view. Familiarity and class similarity are experiential, i.e., how often an object or a class of objects is observed in a particular viewpoint, and therefore, defy a formal analysis. In contrast, the surface geometry of objects varies in a manner that may be captured by differential geometry and topology. Furthermore, because the surface geometry of objects is highly complex, intuition or other ad hoc methods will not generally provide good tools for understanding when an object shifts from one aspect to another (hence the different interpretations of the Biederman and Gerhardstein (1993), and Hayward and Tarr (1997), results). Consequently, it is here that computational techniques may be applied most effectively.

To address the problem of how to define a view, Koenderink and Van Doorn (1979) have suggested a geometric approach for enumerating all topologically distinct views for a given object. This conception of aspects is consistent with our intuitions about the perception of natural objects: as the viewpoint of an object changes, we assign qualitatively different labels, for instance, distinguishing between the 'front' and the 'profile.' It is the formal analysis of this intuition that will be presented next.

3. Qualitative changes in object perception

As an observer moves his or her viewpoint with respect to an object, its appearance will change. From some viewpoints, a small change in viewpoint only leads to a minor variation in appearance whereas at other viewpoints, more significant changes may occur. For example, when driving along a twisty mountain road and rounding a curve, the scene may suddenly change when a new vista comes into view. Those viewpoints where the appearance of the object does not *qualitatively* vary for any infinitesimal motion of the observer are said to be *stable*. In contrast, drastic, qualitative changes in appearance (named *visual events*) only occur at certain *accidental viewpoints*. All of the neighboring, stable viewpoints can be grouped together into regions, and the accidental viewpoints form the borders that delineate the stable regions. A qualitative description of the object's appearance (as mentioned, referred to as an *aspect*) from a viewpoint within each region and the adjacency relationships between regions can be used to define a view-based representation known as an *aspect graph*.

Aspect graphs and their variants have received much attention in the computer vision community and may provide the basis for a representational format for biological object recognition. Building on newly developed mathematical methods (Koenderink & Van Doorn, 1976), Koenderink and Van Doorn (1979) introduced the mathematical basis for aspect graphs. Concurrently, Freeman and Chakravarty (1980) implemented a recognition system based on mulitple views which incorporated many of the essential ideas of aspect graphs. Though the foundations are almost 20-yr-old, only in the past few years have computational algorithms been proposed and implemented for actually constructing the aspect graph from an object model. Bowyer and Dyer (1991) present a survey (see also, Van Effelterre (1994), for an introduction to aspect graphs) and here, we summarize some of the most relevant results from the literature. Most of the related work in

[3] In the computer vision literature, the terms 'aspect' and 'characteristic view' are used somewhat interchangeably. However, in the biological vision community the latter term has a somewhat broader meaning (e.g., preferred or 'canonical' views; Palmer et al., 1981). For the sake of clarity we have used 'aspect' throughout.

computer vision is concerned with constructing aspect graphs from 3D object representations, perhaps produced from a computer aided design (CAD) system. Far less work has addressed how an ambulant observer, such as a human perceiver, might construct an aspect graph by circumnavigating or manipulating an object. Whether starting with a 3D model or actively observing the object, the ultimate aspect graph is dictated by the object geometry; however, the acquisition details will differ drastically. Here, we are only concerned with understanding the relationship between object geometry and the resulting aspect graph representation.

3.1. Assumptions

Before one can arrive at an aspect graph representation, three issues must be considered. First, what is the 3D shape of the object? Is it polyhedral, laminar, completely smooth, piecewise smooth, randomly textured; or, is it rigid, floppy, pliable, compressible or articulated? Nearly all work has considered rigid objects, progressing from 2D objects (Warman, Baugher, & Gualtieri, 1986) to convex polyhedra (Stewman & Bowyer, 1987; Watts, 1987), to general polyhedra (Gigus & Malik, 1990; Plantinga & Dyer, 1990), to simple, curved objects (Kriegman & Ponce, 1990; Chen & Freeman, 1991; Eggert & Bowyer, 1993), to more general curved objects (Petitjean, Ponce, & Kriegman, 1992; Rieger, 1992).

The second, and perhaps most important question for the purposes of this article, is what constitutes a qualitative description of an object's appearance (a *view*) and how might it change with respect to viewpoint? Following Koenderink and Van Doorn, views are defined in the line drawing resulting from image intensity discontinuities, which themselves arise from either surface normal discontinuities (edges or creases) or occlusion boundaries (Nalwa, 1988). Although they are not relevant to the approach presented here, there are also non-geometric image intensity discontinuities which arise from light source or albedo discontinuities. Finally, this geometric approach is not the only possible definition of an aspect graph. Views might also be defined in terms of a qualitative representation of image irradiance, a description of the visible surfaces (as produced by a laser range finder or photometric stereo for instance; Hebert and Kanade, 1985; Ikeuchi and Kanade, 1988), the number and types of features found by some detection process, or the response of a small number of filters (Riesenhuber & Dayan, 1997).

The third question that must be considered is how should image formation be modeled and consequently how should the relationship between the observer and object be parameterized? The relationship between any two rigid bodies (e.g. an observer and object), has six degrees of freedom, but some of these do not affect the qualitative structure of the image. Let us first consider the more general case of *perspective projection*. The image is formed by the intersection with the imaging surface (a plane for a video camera or a curved retina) of visual rays passing through the center of projection. The location of the center of projection determines the qualitative structure of the line drawing while changes in the location and orientation of the imaging surface merely lead to smooth image deformations (a projective transformation in the case of an image plane). The three coordinates of the center of projection sufficiently parameterize the observer–object relationship. When the observer is distant, effects of foreshortening are diminished, and *orthographic projection* adequately models the projection function. The direction from the observer to the object (the *viewing direction*) determines the line drawing, and a viewing direction is readily represented as a point on a sphere (the *view sphere*). Moreover, the qualitative representation is independent of image plane rotation, translation and scaling. For other qualitative representations of a view (e.g., a representation based on image irradiance), additional parameters may be required to model image formation (e.g. the relationship between the object and the light source or sources).

It should also be noted that nearly all aspect graph models do not account for the acuity of the eye or camera, and instead are assumed to perfectly resolve all detail. A consequence is that the number of distinct views can become very large, even though it is unlikely that the differences could ever be resolved. It has generally been presumed that if resolution is directly considered, the resulting aspect graphs would more accurately account for the actual sensor performance and the number of distinct views would become manageable. While Ikeuchi and Kanade (1989) incorporated a sensor model when constructing aspect graphs using a tesselation of the view sphere, they did not consider line drawings. Shimshoni and Ponce (1997) introduced a method for constructing finite resolution aspect graphs of polyhedra. Surprisingly, the number of views increased; the reason is that with a finite resolution sensor with discrete pixels, there are more ways for contours to meet and change their configuration. An alternative approach is to introduce the notion of a scale-space in which the image is blurred to varying degrees prior to extracting the qualitative features used to define a view; an aspect graph structure could then be constructed over the combination of viewpoint and scale (Eggert, Bowyer, Dyer, Christensen, & Goldgof, 1993). Unfortunately, the study of this process is in its infancy, and current approaches make compromises. E.g., in their method for solids of revolution (Pae & Ponce, 1999), blur the object to different degrees (the scale) and then assume that an idealized line drawing is obtained of the blurred object. However, given the

highly limited knowledge in the field of computational vision regarding scale-space aspect graphs, we will not consider this issue further in this paper.

3.2. Aspect graphs and object geometry

As mentioned, we follow Koenderink and Van Doorn and represent an object's image by its line drawing. By way of example, two types of objects will be considered: convex polyhedra and smooth objects. Both perspective and orthographic projection will be discussed.

3.3. Polyhedral objects

A polyhedron is bounded by planar faces, two of which meet at straight edges; three or more faces meet at a vertex. A stable line drawing of a polyhedron is composed of straight line segments that are the image of the edges. The end-points of image segments are either the projection of vertices or t-junctions (where a face occludes an edge). Thus, the qualitative line drawing can be thought of as a graph (sometimes called an *image structure graph*, Malik (1987)) whose labeled nodes are the segment end-points and whose arcs are the straight line segments. Though the visual events, where this graph structure changes, come in two flavors (Gigus & Malik, 1990; Plantinga & Dyer, 1990), for the sake of conciseness, we will only consider the one associated with convex polyhedra. The infinite plane defined by a face divides the space of perspective projection viewpoints in two. From one side of the plane, the entire face including its edges is visible. From the other side, the face itself is occluded, and the visibility of its edges will depend upon the relative location of the viewpoint and the other faces. Thus, as an observer crosses this plane, the visibility of the face changes, and the structure of the line drawing changes, i.e., a visual event. Each face of a convex polyhedron defines a plane, and taken together the set of planes partitions the view space into a number of distinct cells. For every viewpoint within a cell, the line drawings will be qualitatively similar. Note that the qualitative structure is independent of the 3D orientation of the retina.

Fig. 1 depicts an example for a 2D convex polygonal object and a one-dimensional (1D) image; the 2D perspective projection viewpoint space is divided by lines (dashed) defined by the polygon's edges. The views associated with each region can be used to define the nodes of the aspect graph, and the borders between regions define the arcs. Note that some cells have finite area while others extend outward and have infinite area. Orthographic projection can be considered the limiting case of perspective as the viewer becomes infinitely far from the object. The intersection of the visual event planes for perspective projection with a circle (a sphere in three-dimensions) of infinite (or very large) radius partitions the circle into a set of compact regions. These regions define the stable views under orthographic projection. In Fig. 1 there are 15 stable views under perspective projection while there are only 10 stable views under orthographic projection.

For 3D polyhedra, each face defines a plane which divides the 3D viewpoint space into two regions. For convex polyhedra, this is the only type of visual event surface-t-junctions cannot occur in the image of a single convex polyhedron. For more general polyhedra, an additional visual event occurs at those viewpoints where the visibility of t-junctions changes. This set of accidental viewpoints lies on a curved surface in the viewpoint space.[4]

3.3.1. Smooth objects

Let us now consider the line drawing of an object bounded by a smooth surface and its aspect graph. The image contour of a smooth surface arises from the set of surface points (the contour generator) where the line-of-sight grazes the surface. The line-of-sight is ei-

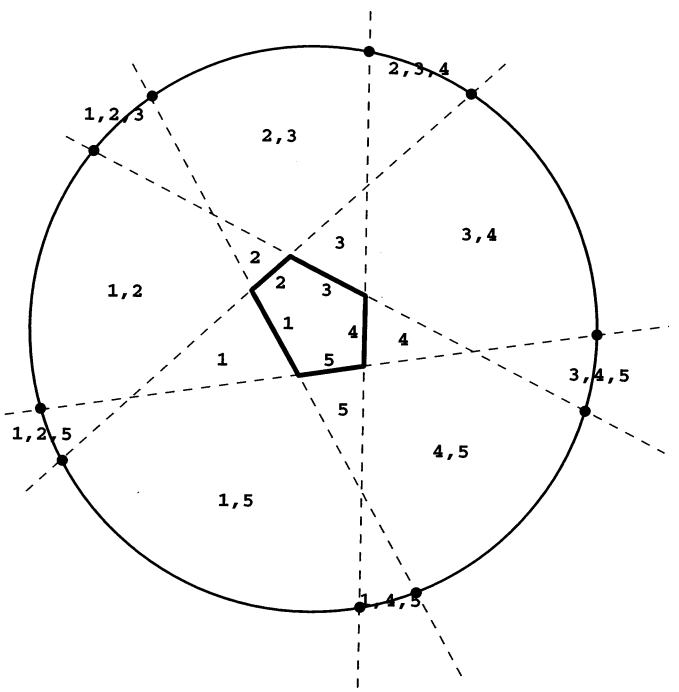


Fig. 1. Perspective projection aspect graphs in two dimensions: The numbered edges of the convex polygon define visual event lines (dashed) which partition the view space into 15 regions. The numbers in each region indicate which of the polygon's edges are visible from all viewpoints within that region. The 'large' circle represents the space of orthographic projection viewpoints, and it is partitioned into 10 arcs by the visual event lines.

[4] For a polyhedron with n faces observed under perspective projection, it has been shown that there are $O(n^2)$ different views when the polyhedron is convex and $O(n^6)$ different aspects when the polyhedron has concavities (Gigus & Malik, 1990; Plantinga & Dyer, 1990).

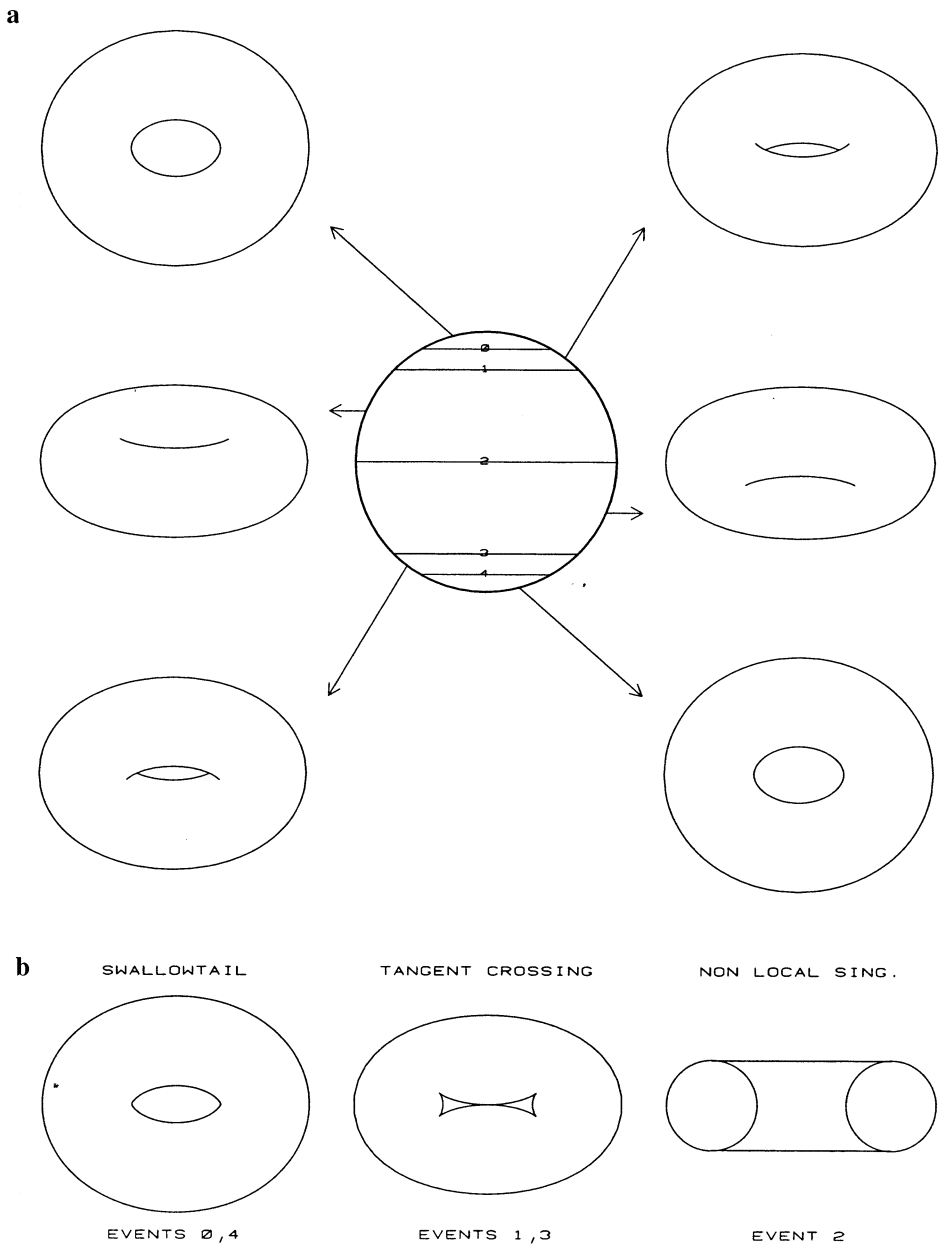


Fig. 2. The aspect graph of a torus. (a) The view sphere is partitioned into six regions corresponding to stable views, or aspects. (b) These regions are separated by five singular views, or visual events, which are instances of the singularities illustrated here (in this and subsequent figures, visual events are designated by numbers and aspects are designated by letters). Reprinted from Kriegman and Ponce (1990).

ther the viewing direction under orthographic projection or a ray emanating from the center of projection under perspective. In contrast with polyhedra whose contour generators (edges) are fixed on the surface, the contour generators of smooth objects depend on the viewpoint and are not rigidly affixed to the surface. From a stable viewpoint, the contour generators (sometimes referred to as the occluding contour, the rim, the apparent contour, or the limb) of a smooth compact surface will be a set of smooth closed curves. The image of the contour generator is also a curve, but it may contain singular points (i.e. sharp corners). Consider for example, the drawing of a torus in the upper right of Fig. 2(a) (this example will be discussed in more detail in 'Implementation and Examples'); the drawing is composed of two closed contours: the one corresponding to the outside of the surface is smooth while the contour corresponding to the hole has singular points. As first shown by Whitney (1955) and discussed in Koenderink and Van Doorn (1976), Koenderink (1984), Nalwa (1988), there are two types of singular points: *cusps* and *t-junctions*. Cusps (contour terminators) arise when the viewing direction is aligned with the tangent to the contour generator; in the image, the

contour terminates at a point. At a t-junction, two distinct points on the contour generator project to the same image point. If the object were not opaque, the image contours would cross; however, because of occlusion, one branch of the cross is hidden, hence the appearance of a 'T'. Thus, the line drawing of a smooth object is composed of smooth curve branches that join at t-junctions and terminate at cusps. Reconsider the line drawing of the torus shown in the upper right of Fig. 2(a); the right and left sides of the hole are occluded, and so the image contour contains two t-junctions and two cusps. Other examples of cusps and t-junctions will be shown below.

The development of the theory of line drawings of smooth objects and their visual events is derived from differential geometry, as well as singularity and catastrophe theory Whitney (1955), Thom (1972), Kergosien (1981), Arnol'd (1984), Rieger (1990). Much of this work assumes that surfaces are transparent; opacity simply prevents certain feature points, sections of the occluding contour, or visual events from being seen. Opacity does not introduce any new features or visual events. It should also be noted that up to occlusion, any point on the surface may be a contour generator. Consider a surface point and a line which just grazes the surface at that point. While continuing to pass through the point, the line can be rotated so that it continues to graze the surface. The rotating line will sweep out a plane called the *tangent plane*. From any viewpoint in the tangent plane, this surface point will project to a point on the image contour.

Thus, we have established the condition for a surface point to project to an image contour; now, what are the conditions for that point to be a singularity (a t-junction or cusp)? To form a t-junction, two points on the surface must project to the same image point. If the line connecting any two surface points lies in the tangent plane of each point, then these two points will project to a t-junction whenever both points lie along a line of site. To understand the conditions for a surface points to project to a cusp, we must consider the differential (local) geometry of the surface. As shown in Fig. 3, a smooth surface can be decomposed into two types of regions, *elliptic* or *hyperbolic* regions which are separated by a *parabolic curve*.[5] While any point on the surface can lie on the occluding contour, cusps are the image of hyperbolic points (locally similar to a mountain pass or a smooth pleat in a piece of cloth). Any hyperbolic point can project to a cusp, but the viewpoint must fall on one of two special lines within the tangent plane called the *asymptotic directions*.

As demonstrated in the above discussion and illustrated in Fig. 2, the line drawing of a smooth surface is composed of smooth curve branches that join at t-junctions and terminate at cusps. How does this structure change for a change of viewpoint? As shown in Kergosien (1981), Arnol'd (1984), Platonova (1984) and Rieger (1990), there is a small catalogue of six types of visual events; three of these are termed *local events* and only involve a single surface point. Similar to a t-junction, the other three *multi-local events* involve more than one surface point. Figs. 4 and 5 graphically illustrate these events. For each event, the top drawing shows a surface and the location of three viewpoints. From each of these viewpoints, the rendered surface is shown below, and the relevant section of the occluding contour is highlighted. The dashed curves denote the occluded sections of contour generator while the solid curves indicate visible sections of the occluding contour. The visual events across these contours are seen in the second viewpoint (the image labeled 'b'). One may visualize the three images as snapshots from an animation obtained by a camera that is moving along a trajectory from viewpoint a to b to c.

Fig. 4 illustrates the local events. In a *swallowtail* transition (Fig. 4(1)), a smooth contour forms a singularity which then breaks apart into two cusps and a t-junction. As shown by example in the two line draw-

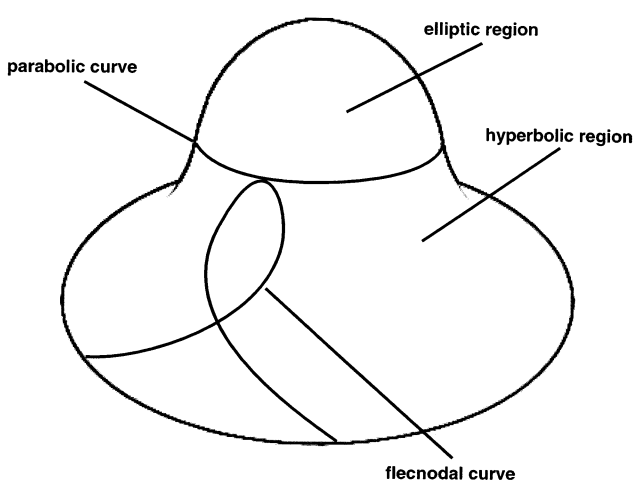


Fig. 3. Points on a smooth surface can be partitioned into elliptic and hyperbolic regions which are separated by parabolic curves. Flecnodal curves lie in the hyperbolic region and may contact the parabolic curve.

[5] At a surface point p, the surface normal is orthogonal to the tangent plane. Consider a plane through p that also contains the surface normal; the intersection of this plane and the surface defines a curve, and the curvature of the curve at p can be determined. Now consider rotating the plane about the surface normal. The normal curvature will vary, and the minimum and maximum values of the curvature are known as the *principal curvatures*. When the principal curvatures have the same sign, the point is said to be *elliptic*. When the signs differ, the point is *hyperbolic*. Since the sign of the normal curvature changes at a hyperbolic point, the curvature must be zero for two directions of the plane; these directions are called the asymptotic directions. At a parabolic point, one of the principal curvatures is zero.

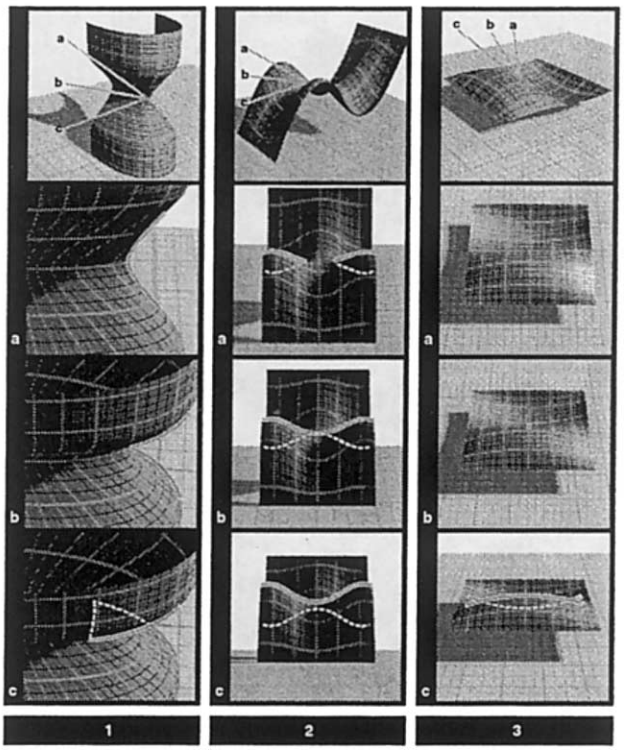


Fig. 4. The change to the image contour during local visual events: (1) swallowtail, (2) beak-to-beak, and (3) lip. The top image shows a view of the surface and the location of three viewpoints while the bottom images depict the surface from those viewpoints. The dashed curves denote occluded portions of the contour generator while the solid curves indicate visible sections of the occluding contour.

ings of a torus in the top of Fig. 2(a), one of the cusps and the two associated branches are occluded when the object is opaque. In a *beak-to-beak* transition (Fig. 4(2)), two cusps approach each other, and the curve branches incident to the two cusps join to form two smooth contours. Finally, in the *lip* transition shown in Fig. 4(3), a closed contour with two cusps appears out of nowhere as the viewpoint approaches the tangent plane of some point on the surface. Only points lying on certain surface curves can participate in a visual event; for each point on the curve, the viewing direction must also be an asymptotic direction.[6] A point on the curve and the asymptotic direction define a line in three space; when taken over the whole curve, these lines sweep out a ruled surface in the viewpoint space. Like the separating planes formed by the faces of a convex polyhedron described in the previous section and shown in Fig. 1, this ruled surface determines the set of accidental viewpoints under perspective.

There are also three multi-local events. These are illustrated in Fig. 5. A *tangent crossing* (Fig. 5(1)) occurs when two distinct surface points project to the same image point as in a t-junction. However, in a tangent crossing, the contour tangents at these two points are aligned instead of intersecting transversally. Imagine for example hiking up a mountain believing that you are heading for the summit when suddenly the true summit emerges from behind the false one; a tangent crossing has just occurred. The transition between the two line drawings in the upper right of Fig. 2(a) is another example. For two points to participate in a tangent crossing, the points must share a common tangent plane. Consequently, a line connecting the two points will lie in this common tangent plane. Pairs of points satisfying this condition form a pair of curves on the surface. The line between corresponding points can be swept along the curves to define the ruled visual event surface. A *cusp crossing* event (Fig. 5(2)) occurs when a point on the occluding contour projects to a

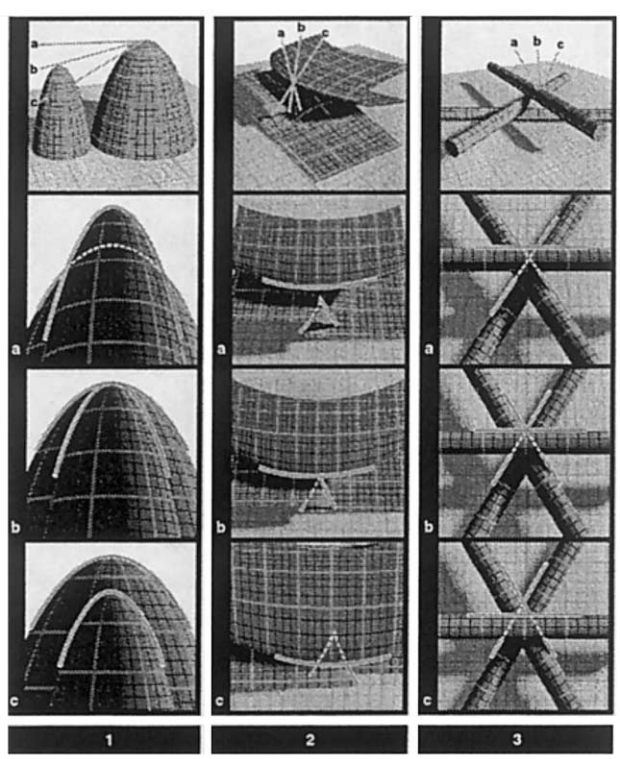


Fig. 5. The change to the image contour during multi-local visual events: (1) tangent crossing, (2) cusp crossing, and (3) triple point. The triple point involves three surface points while the tangent crossing and cusp crossing only involve two.

[6] Points on parabolic curves, which are generically smooth closed curves and separate elliptic from hyperbolic regions as seen in Fig. 3, project to lip and beak-to-beak transitions. Swallowtail transitions occur at the image of points lying on flecnodal curves. As shown in Fig. 3, these curves lie in the hyperbolic region of the surface. They may meet the parabolic curve and be self intersecting. See Koenderink (1990), Petitjean et al. (1992) for a characterization of the differential geometry of flecnodal curves.

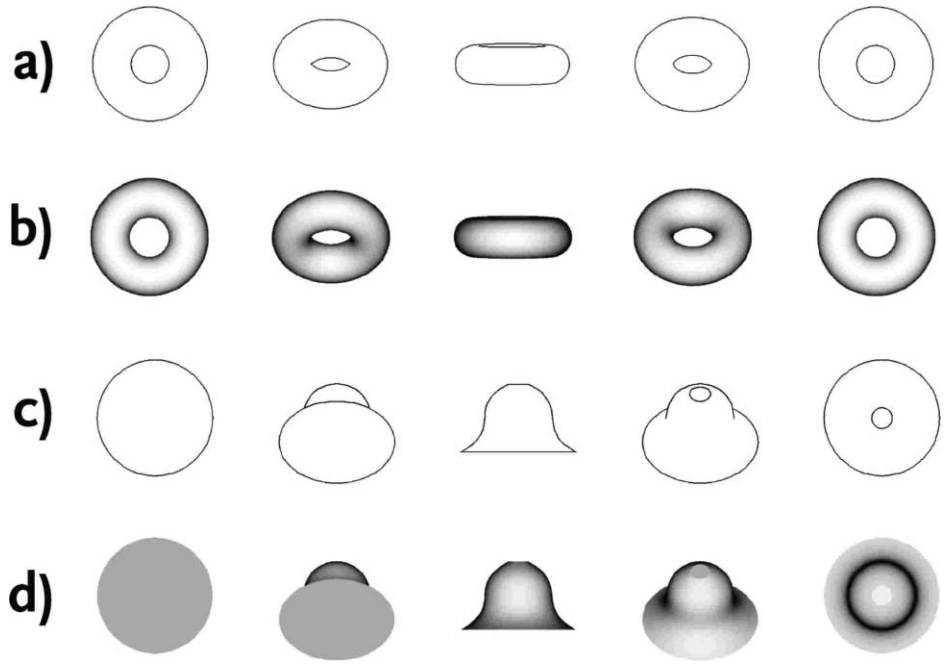


Fig. 6. Examples of stimulus images used in the psychophysical experiments. Each object is displayed at the 45° increments. (a) Line drawings of the torus, (b) shaded images of the torus, and (c) line drawings of the bell. (d) Shaded images of the bell. Note that these are the actual line drawings used in the experiments. Because of the rendering method the visible contours are bitmapped as opposed to vector-based, hence their somewhat jagged appearance.

cusp and another point on the occluding contour projects to the same image point. A *triple point* occurs when three distinct points on the contour generator project to the same image point; as seen in Fig. 5(3), the arrangement of t-junctions changes across the visual event. As in the tangent crossing, the pair or triplet of surface points participating in these latter two events define a line; taken over the whole object, the locus of such points generates a ruled surface that partitions the viewpoint space.

Between the local and multi-local events, a set of ruled visual event surfaces are associated with the object. These ruled surfaces partition the viewpoint space into regions, and within each region the line drawing will be qualitatively stable since all possible changes only occur on these surfaces. That is, for any viewer motion within the region, no new singular points (cusps or t-junctions) will form or disappear, and their interconnectivity will remain unchanged. Thus, all qualitatively distinct line drawings can be enumerated with one per region. As in the case of convex polyhedra, the intersection of an infinitely large sphere with these visual event surfaces partitions the sphere into regions. Each region corresponds to a stable view in the orthographic projection aspect graph. In addition, the area of a region is proportional to the probability that a particular view will be seen for a random, uniformly distributed orthographic projection viewpoint.

3.4. Implementation and examples

The above description of the relationship of object geometry to the visual events forms the basis for an implemented algorithm for constructing the aspect graph of objects modeled by algebraic surfaces. Specific details of the algorithm, a set of equations characterizing the visual events, and examples under orthographic projection can be found in Petitjean et al. (1992). See Rieger (1992) for an alternative algorithm. It should also be mentioned that for piecewise smooth objects, the catalogue of stable singularities (Malik, 1987) and visual events (Rieger, 1987; Sripradisvarakul & Jain, 1989) includes the ones for smooth surfaces as well as some additional types of events.

Two objects, which are surfaces of revolution, were used in the psychophysical experiments described in the following section and illustrated in Fig. 6. For both objects their orthographic projection aspect graphs have been computed (Kriegman & Ponce, 1990) as shown in Figs. 2 and 7. Because of the axis of symmetry of a surface of revolution, the partitioning of the view space will also be circularly symmetric about the axis (i.e. rotating the viewpoint about the axis leads to the same line drawing). Under orthographic projection, the view sphere is partitioned along lines of latitude which can be described by a single number. This both simplifies aspect graph construction (Kriegman & Ponce,

1990; Eggert & Bowyer, 1993) and simplifies the interpretation of psychophysical experiments; when probing participants, we are drawing sample images from a 1D space rather than a higher dimensional one and do not have to impose an ad hoc metric for the 'distance' between images.

The torus shown in Fig. 2 is about the simplest smooth object that has a non-trivial aspect graph. Even so, there are six distinct views, separated by five visual events. Because of the object's bilateral symmetry, the aspect graph also exhibits bilateral symmetry about the equator. As the viewpoint moves from the north pole (looking along the axis of revolution), the line drawing is composed of two concentric smooth curves. At some viewpoint, a double swallowtail transition occurs, and two cusps and two t-junctions come into view. As the viewing direction approaches the equator, a tangent crossing occurs, and the entire hole is no longer visible. The rearward contour generator of the hole has become completely occluded. When crossing the equator, the leading edge of the hole on the lower side of the torus becomes visible, and the series of stable views repeats in reverse. Fig. 2(b) illustrates the line drawings from each of the accidental viewpoints for a transparent torus.

Fig. 7 illustrates the aspect graph of a bell-like object generated by sweeping a cubic curve around an axis (Kriegman & Ponce, 1990). This object is piecewise smooth, and so the catalogue of visual events is more extensive than described above (Rieger, 1987; Sripradisvarakul & Jain, 1989). The object does not have a symmetry plane orthogonal to the axis, and this is reflected in the aspect graph. One may note that some of the visual events are subtle and somewhat difficult to discern. For example, the qualitative changes between views c and d: the image of the lower edge and the occluding contour form a t-junction in view c, and they share a common tangent in view d (a curvature-l junction). The perceptual salience of such changes will be discussed in the context of the experimental results.

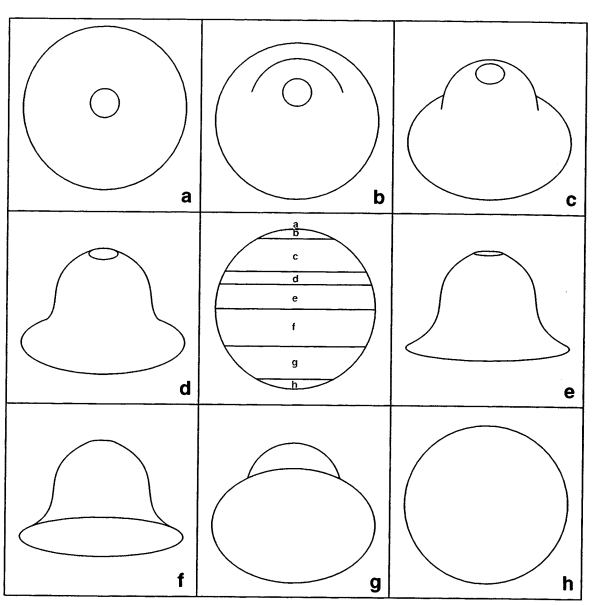


Fig. 7. The partitioning of the view sphere and stable views of a bell under orthographic projection. Adapted from Kriegman and Ponce (1990).

4. View-based representations in human perception

To this point, we have reviewed evidence suggesting a role for view-based representations in human object recognition and argued that theories of view-based representation require a more precise definition of what defines a view. In particular, while it seems clear that familiarity will play some part in determining which viewpoints of an object are retained as mental representations, there are both empirical and computational reasons to suppose that object geometry is also a factor. To this end, we have offered a framework in which geometry is used to partition the complete set of viewpoints of an object into qualitative views. This provides an avenue for understanding what constitutes a qualitative view in human perception. Furthermore, the definition of a view in an aspect graph representation arises from the same physical properties of objects with which our visual systems must contend. Thus, the essential viewpoint-dependent features underlying the aspect graph representation are likely to be fundamental to any view-based representation, including those generally unrelated to the aspect graph approach.

Despite this potential, these claims should be tempered with several caveats concerning the assumptions underlying aspect graphs. In particular, these are assumptions that are unlikely to be true for human observers. First and foremost, the aspect graph representation is structured around *purely* qualitative views that do not capture the perceptually salient *quantitative* variations in object appearance that arise from changes in viewpoint. Empirical evidence indicates that an exclusively qualitative multiple-views representation is inadequate for human perception (see the discussion below). For example, humans have the ability to perceive and discriminate quantitative changes (smooth changes that do not result in qualitative differences) and represent views differing only in quantitative factors (Perrett & Harries, 1988; Tarr & Pinker, 1989; Tarr, 1995). Second, as mentioned previously, most current algorithms for computing aspect graph representations rely on the availability of 3D object models (such as CAD models), rather than being acquired from repeated exposure to viewpoint-specific 2D images over time. Third, aspect graphs are generally computed at infinite resolution and give equal weight to all pertinent images features found in the line drawing of the object; therefore, for even a moderately complex object there

may be a huge number of qualitative views (Kriegman & Ponce, 1990). Even the aspect graphs developed from sensor models with finite resolution retain a large number of views (Shimshoni & Ponce, 1997). The potential for a huge number of views per an object raises the concern that, as they now stand, aspect graph representations will be too complex to be parsimoniously represented or utilized in human perception.

Thus, while aspect graphs provide an attractive framework for understanding view-based representations in humans, it is important to acknowledge that their actualization will be necessarily somewhat different from the approach reviewed in the preceding section. However, this does not automatically render current work on aspect graphs irrelevant. Indeed, our claim is that the image features that define visual events may be helpful in understanding 3D object geometry for human perception. Specifically, the image features that define visual events in the computational theory may be one of the factors used by humans to organize viewpoints of objects during the acquisition of view-based object representations. Moreover, methods for the geometric partitioning of viewpoints of 3D objects are necessary for interpreting patterns of viewpoint-dependent behavior obtained in empirical studies of recognition. Thus, formal descriptions of object geometry, including but not limited to current aspect graph methods, offer a principled means for analyzing human recognition performance and perceptual behavior, as well as a model for how to manipulate viewpoint with regard to stimulus object geometry.

4.1. Psychophysical studies

In order to assess the viability of this framework, we present several psychophysical studies. We investigated whether humans are indeed sensitive to the image features used to characterize the topologically distinct views of an aspect graph as defined in the previous section. The essential idea of these experiments is that as an object's orientation changes, humans perceive qualitative, as well as quantitative (e.g., shape), changes in the appearance of an object. Therefore, judgments about the orientation of an object should not be uniform; rather, for orientation pairs that are qualitatively similar, discriminating between them should be difficult, but for orientation pairs that are qualitatively different, discrimination should be much easier. Specifically, we used the aspect graph decompositions shown in Figs. 2 and 7 of two smoothly curved objects (Kriegman & Ponce, 1990), a torus and a bell, as models to predict the orientations where participant performance was expected to be good and where it was expected to be poor. In judging whether two images of an object are at the same or different orientations, participants' accuracy was expected to be higher for orientation pairs that span a visual event as predicted by the computational theory. In such instances we may infer that humans are sensitive to the particular configurations of features that define that visual event.

4.1.1. Methods

Participants: Participants were primarily drawn from the undergraduate introductory psychology course at Yale University (New Haven, CT) and were provided with credit for their participation. A number of additional participants were paid 5 dollars for their time. The number of participants for the five experiments was 24, 33, 26, 27, and 28 respectively. Participants were not used in more than one experiment and all were naive as to the purposes of the study.

Stimulus materials: Two solids of revolution, a torus and a capped bell as illustrated in Fig. 6, were used as stimuli. Both depict smoothly curved objects similar in appearance to objects often encountered in the natural world (in contrast to the stimulus objects used by Shepard and Cooper (1982), Rock and Di Vita (1987), Corballis (1988), Bülthoff and Edelman (1992) and Tarr (1995)). Because various computer graphics techniques are available for depicting object models from a specified vantage point, it is important that the chosen method does not bias participants' responses. For example, if the object is treated as being glossy, the shape and location of specularities could provide cues about light source location; this might aid in determining orientation differences. Here, we have used two depiction methods: line drawings and shaded images. To render a line drawing, there is no need to introduce light sources that could confound experimental results. However, line drawings themselves result from applying an ideal edge detection process; therefore, the resultant line drawing may be somewhat different from that produced by early visual processing in humans. Reinforcing this point, a recent study found that images of objects generated from the output of edge detectors are identified much more poorly than idealized line drawings or shaded color images (Sanocki, Bowyer, Heath, & Sarkar, 1998). Furthermore, as participants in this experiment, observers' edge or line detection processes will be applied to the line drawing stimuli, and the output that will result is likely to differ from what would occur if an intensity image were directly observed. Consequently, stimulus objects have also been rendered as shaded images of a Lambertian surface with a single light source in the same direction as the viewing direction. Note that ideal line drawings of such shaded objects (corresponding to intensity discontinuities) would be identical to those produced by the first depiction method. No additional intensity discontinuities due to shadows, specular reflections, or pigmentation changes are introduced by this choice of lighting direction and rendering model.

Design and procedure: The participants' task was to judge whether two consecutively presented images of the same smoothly curved object were displayed at the same or at different orientations (or equivalently, observed from the same or from different viewpoints). A trial consisted of an object displayed at a 'target' orientation for 250 ms, followed by the same object in a 'probe' orientation for 250 ms; both the target and the probe were masked by a gray field displayed for 100 ms.[7] Changes in viewpoint were produced by a depth rotation around a horizontal axis parallel to the image plane where the major axes of the object were initially aligned with the gravitational and screen defined major axes. Participants responded by pressing one of two keys: the 'S' key for same responses and the 'D' key for different responses; participants were given feedback in the form of a beep when their response was incorrect. Participants' viewing positions were fixed approximately 54 cm from the screen by the use of a chin rest: this resulted in the upright viewpoint of the torus and the bell, respectively, subtending $7.95° \times 7.95°$ and $9.53° \times 9.53°$ regions of visual angle.

In a given experiment, the same object was used throughout 720 randomly order trials. The object appeared 20 times each at the 36 target orientations defined by rotations from 0° to 180° in 5° increments (where 0° is the upright position). Experiment 2, which used the shaded torus as the stimulus object, differed in that it included 5° increments starting from 2.5° rather than 0° to ensure that no single image fell directly on a visual event). For each target orientation, the orientation difference from probe to target was either 0°, 5°, 10°, or 15°, with the proportion of trials for each difference being 40%, 20%, 20%, and 20%, respectively. In addition, differences of 5°, 10°, and 15° were split evenly between forwards and backwards rotations. Experimental sessions were divided into two identical 360 trial blocks to allow participants some respite. Each block was preceded by four practice trials that were not included in any of the analyses.

Five distinct experiments were run. Experiments 1 and 2 used the torus as the stimulus object, while Experiments 3–5 used the bell as the stimulus object. Experiments 1 and 3 used objects rendered with only occluding contours and edges, while Experiments 2, 4, and 5 used objects rendered with smooth shading. Other than the variation in the particular object used, Experiments 1–4 were identical in that the position of the object on the screen remained fixed, while Experiment 5 added a random position shift from the target to the probe of between -50 and $+50$ pixels in both the horizontal and vertical directions. This manipulation was introduced to ensure that participants did not simply fixate on a local region of the screen that they believed to be diagnostic for performing the orientation discrimination.

4.1.2. Results and discussion

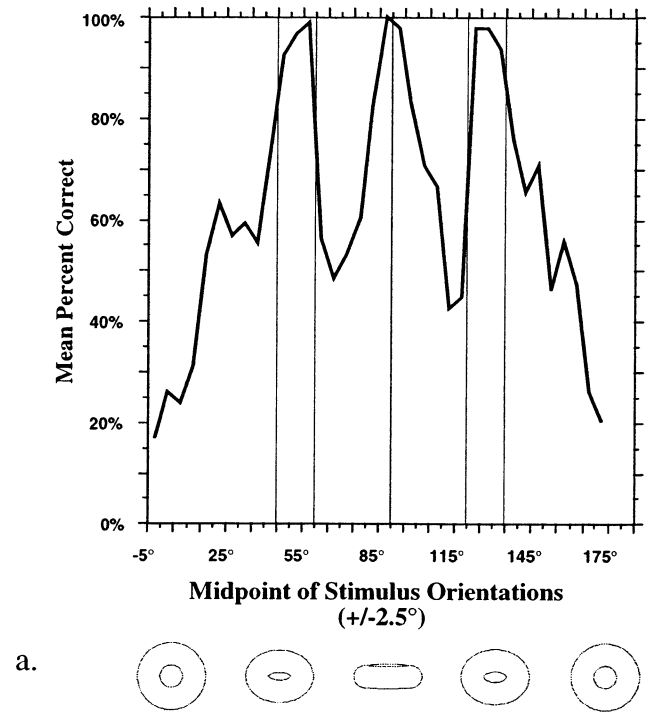
Performance was assessed by measuring participants' accuracy in detecting an orientation difference between the two images. If a participant failed to respond within 7500 ms the trial was considered incorrect. Orientation sensitivity functions were computed separately for target–probe separations of 5°, 10°, and 15°;[8] in each instance, data from forwards and backwards rotations were combined according to the midpoint of the probe and target orientations. For example, view orientation pairs of 15°/20° and 20°/15° (both 5° target–probe *separations*) are included in the mean for 17.5°. Orientation sensitivity functions for 10° and 15° target–probe separations were uniformly at ceiling with almost perfect performance at all orientations[9]. Therefore, the following analyses focus only on 5° separations.

General results: Figs. 8 and 9 illustrate participants' mean percent correct in discriminating views of given stimulus object. In each instance, measured accuracy reflects participants' ability to discriminate two views of the object separated by 5° of rotation in depth. Each data point represents an orientation region $\pm 2.5°$ around the specified midpoint. For instance, the value

[7] The interval between the target and the probe was 100 ms (during which the first mask was presented). This interval is within the range found by Ellis and Allport (1986) to result in reliable effects of object viewpoint on performance. In contrast, Ellis and Ailport found that longer intervals, e.g., greater than 750 ms, reduced the effect of viewpoint on recognition performance. The shorter interval was employed because we were specifically interested in how different changes in viewpoint influence object memory. Of course, the shorter interstimulus interval opens up the possibility that participants are responding based on local image features, an explanation ruled out by the results of Experiment 5.

[8] Trials where the probe and target were identical were not predicted to yield meaningful patterns of responses due to the absence of any changes in visible features between the two images. This was found to be the case at all viewpoints participants' accuracy was generally higher than that found for different trials, ranging between 60% and 95%.

[9] That participants were at ceiling at the larger angular separations does not restrict the generality of qualitative changes to only small rotations. Experiments often manipulate stimulus presentation through duration, degradation or masking to 'probe' the normally opaque processing of the stimulus. Specifically, ceiling effects do not reduce the importance of qualitative changes. In assessing how views are delineated, quantitative changes, such as those resulting in the at-ceiling performance, do not provide any decomposition of the view sphere into unique views. In contrast, qualitative changes, regardless of angular separation, will group some regions of the view sphere as similar and others as dissimilar. If perception relied on quantitative information, there would be no principled method for determining when a change was dramatic enough to warrant a new view or trivial enough to be considered a familiar view. Thus, every slight quantitative change (e.g., every change in viewpoint) would lead to a new unique view; clearly an unparsimonious model.



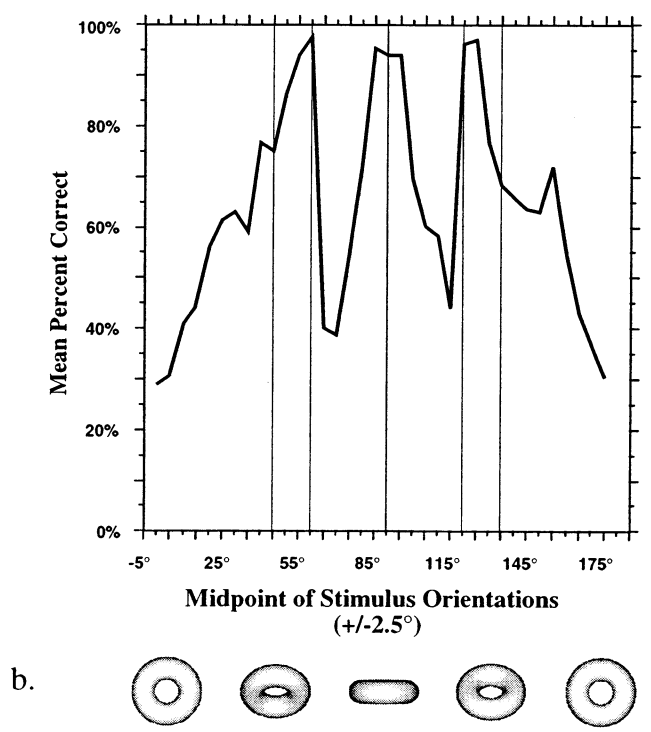


Fig. 8. Mean percent correct in discriminating views of a torus. Measured accuracy of participants' ability to discriminate two views of a torus separated by 5° of rotation in depth. Each data point represents an orientation region $\pm 2.5°$ around the specified midpoint. For instance, the value at 87.5° denotes accuracy for differentiating the torus when displayed as the target–probe pairs 85°/90° and 90°/85°. Locations of the visual events predicted by the formal theory are marked by the vertical gray lines. (a) Experiment 1. Percent correct for discriminating views of a line drawing of the torus. (b) Experiment 2. Percent correct for discriminating views of the shaded torus.

at 87.5° denotes mean percent correct for differentiating an object when displayed as the target–probe pairs 85°/90° and 90°/85°. The interpretation of such data is relatively straightforward: the higher accuracy score, the better the participants were at discriminating the two orientations defining the midpoint indicating that they are more sensitive to the changes in image features between these orientations.

Locations of the visual events predicted by the formal theory are marked by the vertical gray lines according to the stimulus object used in each experiment. The crucial characteristic of each function is the prominent maxima; that is, the orientations where participants' accuracy is greatest relative to not only the surrounding orientations, but to the overall mean accuracy across all orientations. When one compares these maxima to the predicted accidental viewpoints (the orientations where the aspect graph makes the transition from one view to another), accuracy in discriminating orientations does increase when images cross a visual event. Note that because we made no predictions about the pattern of responses for orientations *within a single view*, we cannot apply inferential statistics to the analysis of these data. Suffice it to say, that, in general, when maxima do occur, they do so at orientations predicted by the formal theory. We now turn to the discussion of the results from individual experiments.

Experiment 1: As illustrated in Fig. 8(a), there is a systematic change in performance with changes in the orientation of the torus. Specifically, at all five of the visual events (actually three different types of transitions due to symmetry) participants' percent correct in detecting the orientation difference is highest at target–probe pairs that span visual events occurring at 45°, 60°, 90°, 120°, and 135°. At first pass, these results suggest that observers are sensitive to particular classes of transitions in image features within objects rendered with edges and occluding contours.

A second issue concerns the pattern of responses at orientations between visual events. Here, we observe a systematic increase in performance as the orientation pair approaches a visual event. The visual events present for the torus all correspond to transitions in the configurations of images features located near the left and right edges of its inner rim. However, other changes, albeit non-qualitative, occur in the surrounding silhouette, both in terms of its shape and the total area it defines. Similar non-qualitative changes occur in the inner rim. Such changes are not restricted to visual events, but rather are distributed throughout all orientations, with more dramatic changes coincidentally occurring at orientations approaching visual events (e.g., a 5° rotation while the torus is viewed near on-end will produce a greater change in area than will a 5° rotation when the torus is viewed from the side). Therefore, it is probable that participants are able to use these relative

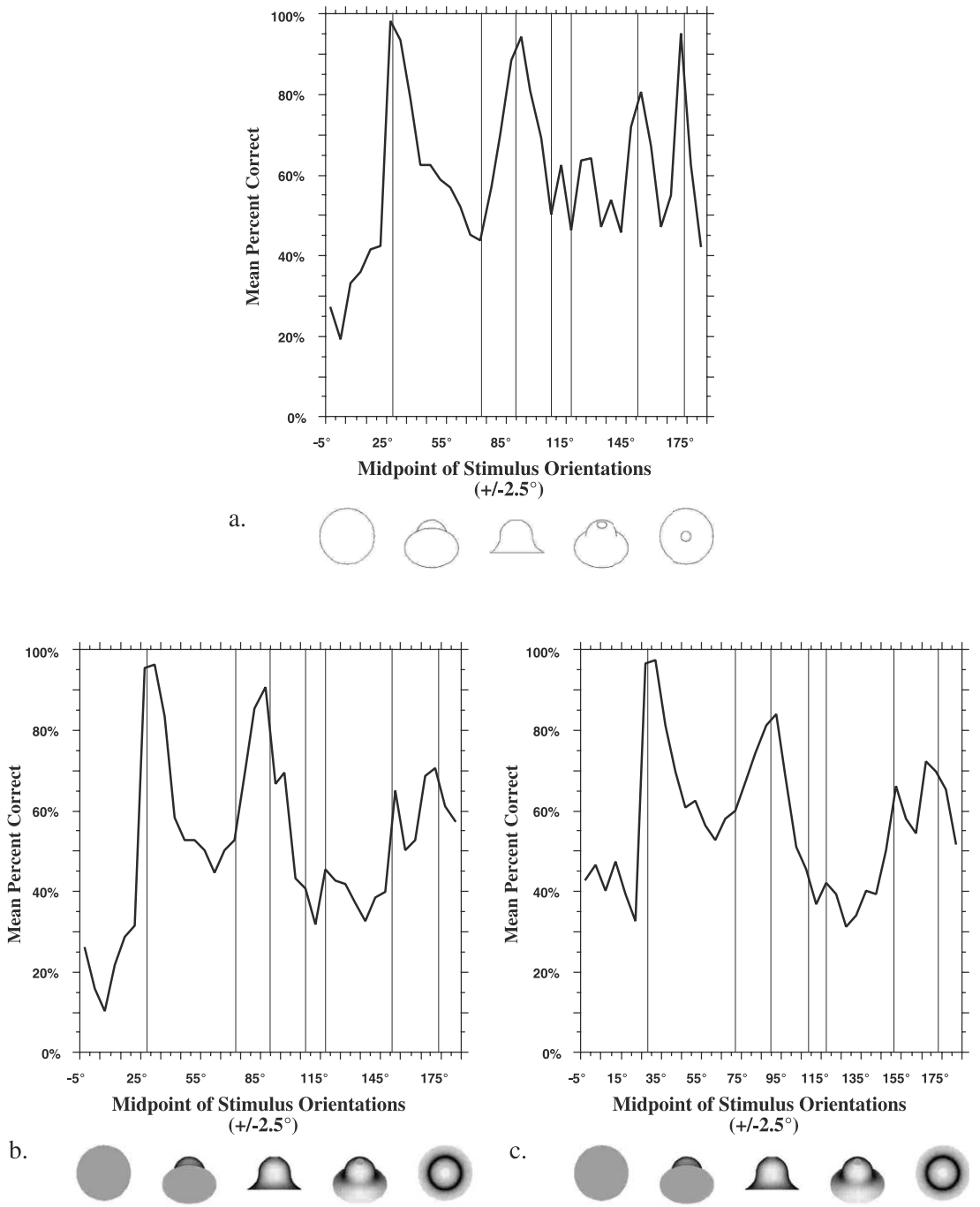


Fig. 9. Mean percent correct in discriminating views of a bell. Measured accuracy of participants' ability to discriminate two views of a bell separated by 5° of rotation in depth. Each data point represents an orientation region $\pm 2.5°$ around the specified midpoint. Locations of the visual events predicted by the formal theory are marked by the vertical gray lines. (a) Experiment 3. Percent correct for discriminating views of a line drawing of the bell. (b) Experiment 4. Percent correct for discriminating views of the shaded bell. (c) Experiment 5. Percent correct for discriminating views of the shaded bell with a shift in position (a random ± 50 pixel shift in both the horizontal and vertical directions the resolution of the monitor was 72 pixels per inch, resulting in shifts of approximately 2.5° of visual angle in each direction).

changes in shape and area to enhance their estimates of orientation, thereby exhibiting differential performance for within-view orientations. While not entirely surprising, this result is at odds with models in which participants are judging orientation based *solely* on qualitative image structure (which would predict uniformly poor performance at such orientations). Consequently, performance theories of how humans encode viewpoint-dependent information in object representations should ultimately include parameters beyond those needed for computing aspects (for instance, see Perrett and Harries (1988)).

Experiment 2: As illustrated in Fig. 8(b), a similar pattern may be observed for the smoothly shaded torus. Again, performance is highest at target–probe pairs that span the visual events predicted by the computational theory. However, in contrast to the results of Experiment 1, maxima are observed at only three midpoints, the visual events at 60°, 90°, and 120°. Apparently, in shaded images, the swallowtail transitions (from a smooth curve to a t-junction and a cusp) located at 45° and 135° are not sufficiently salient to produce qualitative differences among orientations. Indeed, while accuracy at these two orientations is higher than at preceding or subsequent orientations, it is only by virtue of the overall trend towards better performance as orientation approaches salient visual events. This finding indicates some of the limits on the types of image features and transitions that play a role in human object perception. Moreover, the results of this experiment are significant in that not only do they confirm some of the results obtained for line drawings of the same stimulus object, but they do so using smoothly shaded images that are similar in appearance to objects found in the 'real-world,' e.g. rendered in a manner that does not introduce biases about the saliency of particular features. While a perfect edge detection process would produce the line drawings used in Experiment 1, the human visual system computes edges using receptive fields; therefore, the shaded images used in Experiment 2 will not yield edge maps identical to the 'ideal' case. Taken together with the results of Experiment 1, these findings confirm that humans are sensitive to some of the features used in one approach for defining aspects.

Experiments 3–5: As illustrated in Fig. 9, the use of a complex piecewise smooth object, in this instance the capped bell, yields a somewhat more complex pattern of responses than that obtained in Experiments 1 and 2. Specifically, as shown in Fig. 9(a) prominent maxima are only observed at four of the predicted visual events: 28.8°, 90°, 151.2°, and 174.3°, although crucially, as in the previous experiments, when maxima are found, they occur at computationally predicted orientations and never at orientations for which no prediction of improved performance was made. On the other hand, at the other three visual events, located at 72.2°, 107.8°, and 117.8°, relatively low accuracy is actually found. Remarkably similar patterns are found in Experiments 4 and 5, adding, respectively, smoothly shaded images (Fig. 9(b)) and random shifts in the relative positions of the target and probe (Fig. 9(c)). Thus, we can generally conclude that participants are sensitive to the changes found at certain visual events and insensitive to the changes found at others regardless of the rendering technique used for display. Additionally, the similarity of the results of Experiment 5 to Experiments 3 and 4 indicates that participants are processing the intrinsic relationship of image features, rather than simply focusing on a fixed region of the display that is diagnostic for the task.

In light of these results, one intriguing question is why participants were completely insensitive to certain visual events. First, let us examine the exact nature of the three 'ignored' visual events. At 72.2° there is a transition from a pair of features (a t-junction and cusp) to a 3-tangent junction (Fig. 10); at 107.8° there is a transition from a curvature l-junction to a t-junction and cusp; and, at 117.8° there is a transition as a limb (the backside of the bell) becomes disoccluded (Fig. 11).

Why might observers ignore these particular events? First, particularly for the event at 72.2° (Fig. 10), the changes in the configurations of image features are quite subtle. Of course, this inference is tautological in that such transitions most likely seem subtle to us and to our participants for exactly the same reasons. Therefore, while experimental results confirm our intuition, it could not have been predicted by the present computational theory.

Second, particularly for the events at 107.8° and 117.8° (Fig. 11), changes occur only gradually over shifts in orientation. This can be seen in the leftmost and rightmost images in Fig. 11, where over a wider range of orientations, the transitions in both configurations of features are clearly visible and quite salient. This is consistent with the hypothesis that features at some scales are ignored, thereby significantly reducing the complexity of many perceptual processes and, crucially, the number of aspects per an object. This claim is at odds with most computational derivations of aspect graphs, except (Eggert et al., 1993; Shimshoni & Ponce, 1997), since they ignore scale. This assumption, however, results in a huge number of views – one

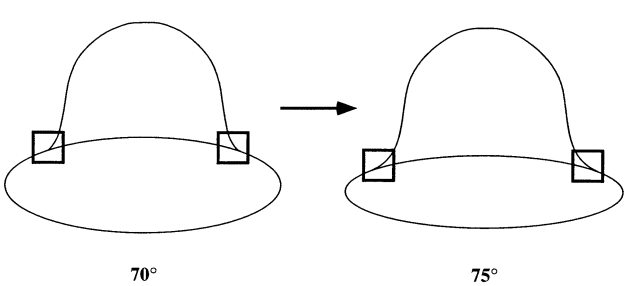


Fig. 10. Views of a bell on either side of visual event at 72.2°. Within the boxed region of the left view, a *t-junction* is formed by the upper contour joining the smooth lower contour at a non-zero intersection angle; in the right view the intersection is tangential and forms a *three-tangent junction*. This illustrates an example of a visual event predicted by the computational theory, but where psychophysical results indicate that observers are not sensitive to the particular configuration of image features. Such findings may be useful in developing parsimonious object representations and practical applications using aspect graphs.

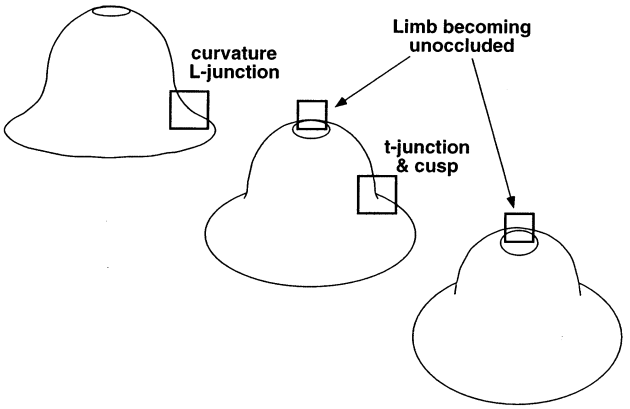


Fig. 11. Views of a bell across visual events at 107.8° and 117.8°. Within the boxed region in the lower right of the left view, a *curvature l-junction* is formed by a tangential intersection between two contours, in the center view the two join at a non-zero intersection angle, producing a *t-junction and cusp*; the boxed region near the top of the center view also illustrates an occluded limb (the backside of the bell), in the right view this limb has become disoccluded. While current psychophysical results indicate that observers are not sensitive to these visual events, there may be conditions under which they become more salient; for instance, by varying their relative scale relative to the overall image. Such findings may be useful in developing finite-scale aspect graphs, as well as understanding how humans encode features within object representations.

ubiquitous criticism of the aspect graph approach. Moreover, if aspect graph-like structures are to be used by active observers that acquire information over finite amounts of time, e.g., humans, then the number of potentially different views per an object must be reduced (such objectives are also attractive for reasons of general parsimony). Therefore, a better understanding of the mechanisms by which humans encode viewpoint-dependent information, in particular in terms of partitioning the aspects of objects, will hopefully aide in the development of finite-scale aspect graph construction algorithms, as well as place constraints on the features within an image that are considered relevant or salient to a particular perceptual task.

5. General discussion

We began this article by presenting evidence that human object recognition is mediated at least in part by view-based mechanisms and representations. This claim is supported by a growing body of psychophysical (Bülthoff & Edelman, 1992; Biederman & Gerhardstein, 1993; Tarr, 1995; Hayward & Tarr, 1997; Tarr, Williams, Hayward, & Gauthier, 1998) and neuroscientific (Warrington & Taylor, 1973; Layman & Greene, 1988; Perrett et al., 1989; Plaut & Farah, 1990; Logothetis & Pauls, 1995) results indicating that humans and other primates mentally represent 3D objects as sets of views, each view corresponding to a more or less limited range of adjacent viewpoints. While this approach may offer some computational advantages over the more standard 3D reconstruction approach (Marr & Nishihara, 1978), it also introduces a host of new computational issues, the most salient of which is defining what constitutes a 'view.'

One approach to the question of 'What defines a view?' is primarily empirical. It is well established that humans are sensitive to the frequency of occurrence of a given object in a given orientation (Tarr & Pinker, 1989; Tarr, 1995) and to visually-similar objects in a given orientation (Jolicoeur & Milliken, 1989; Moses et al., 1996; Gauthier & Tarr, 1997b; Tarr & Gauthier, 1998). Building on such findings, several groups have investigated how experience shapes both the acquisition and representation of object views. For example, Perrett and colleagues (Perrett & Harries, 1988; Perrett, Harries, & Looker, 1992) used an 'inspection' methodology to examine how observers distribute their time across different viewpoints when learning about new, never-before-seen objects. In one study they attempted to characterize inspection time distributions for relatively simple faceted and smooth objects ('tetrahedra' and 'potatoes') in terms of the image properties of the preferred views. For the faceted objects, the preferred views typically showed a vertically symmetric image centered on an edge or object face; for smooth objects, the preferred views had the major axis of each object aligned or perpendicular with the observers' line-of-sight. Interestingly, in both stimulus conditions the views preferentially inspected by observers did not correspond to aspects as defined by the aspect graph approach. This result was reinforced in a second inspection study using a 'machined tool part' (which unlike the previous study, was composed of multiple distinct parts). Perrett (Perrett et al., 1992) again found that preferred views were determined more by alignment with the line-of-sight and gravitational axes, than by the intrinsic topology of the object (similar results using familiar common objects are reported by Blanz, Tarr, and Bülthoff (1999)). Although these results appear inconsistent with our present results, it is difficult to compare findings between different methodologies. Inspection times during the acquisition of novel objects may require different information than that used to discriminate between views. In particular, when learning about a new object, observers may select the most informative, readily comprehended views, not those that are most stable or specifically distinct from other views. However, the mechanisms whereby one *learns* about a 3D object are not necessarily the same as the mechanisms used to *organize* its representation or *perceive* its 3D structure.

Supporting this interpretation, several studies by Todd, Koenderink and colleagues indicate that topological features of objects play an important role in object perception and memory. For example, Norman and Todd (1994) and Norman, Todd, and Phillips (1995) found that observers were better able to perceive the rigid 3D structure of moving objects when its deforming occlusion contours rotate through multiple aspects. In contrast, when only a single aspect is shown, the objects appear non-rigid as they rotate in depth. Phillips, Todd, Koenderink, and Kappers (1997) had observers identify corresponding points on a single object viewed from more than one orientation. They suggest that the localization of points on smoothly curved surfaces relies on the availability of stable features, such as the minima and maxima of curvature. Finally, in a match-to-sample task Todd, Chen, and Norman (1998) found that observers performed most accurately and responded most quickly when objects could be discriminated based on topological differences as opposed to affine or euclidean differences. Thus, there is some evidence that mechanisms of object perception and recognition are sensitive to topological surface features. Our results reinforce this point and test the importance of two specific classes of features as defined by a formal theory of object geometry. We believe that this approach offers a principled basis for understanding human performance.

To begin address this complex issue, we have reduced the question of 'What defines a view?' to the simpler question of whether observers are sensitive to the kinds of viewpoint-dependent image features used to delineate views in a specific version of aspect graphs. Results from experiments using two smoothly curved objects indicate that for certain types of features human perceivers are better able to perform orientation judgments when the viewpoint crosses a visual event as compared to when the viewpoint does not cross a visual event (as defined by the theory). It is, however, important to emphasize that any positive results in this regard *do not* imply that observers are employing or constructing aspect graphs in the manner proposed by Koenderink or implemented by Kriegman and Ponce. Indeed, as discussed earlier, the conditions under which we normally learn about objects, disparate views over space and time, will necessarily render human object representations somewhat different from those commonly used in computer vision (but see Seibert & Waxman (1990, 1992)). Consequently, the fact that we find that observers are sensitive to some viewpoint-dependent image features suggests only that these features are available as candidates for structuring more complex representations. Given the central role that the field has ascribed to such higher-level object representations, we will now consider some of the possibilities in this regard.

5.1. The role of viewpoint-dependent features in visual cognition

Theories of visual cognition, including both mental imagery and object recognition, are predicated on the assumption that we encode object representations in a visio-spatial format. Furthermore, the organization of such representations is not arbitrary, but rather is hierarchical in that it captures both our exemplar-specific and our categorical knowledge about objects. Hence, an essential component of the recognition process is that we are able to identify both the *particular* instance of an object and its various category memberships, e.g., knowing that an object is a 1965 Mustang Convertible, a Ford, a sports car, and a car; all presumably accomplished through matches between descriptions of the input shape and stored object representations. Current thinking (Jolicoeur, 1990; Tarr & Pinker, 1990; Farah, 1992; Hummel & Stankiewicz, 1996b) suggests that there are at least two variants for object representations: one that is 'image-based,' strongly configural (in that many of these models assume that the specific locations of every feature are known) and highly specific to particular viewpoints (Edelman, 1995; Tarr & Bülthoff, 1995; Riesenhuber & Poggio, 1999), and one that consists of features encoded in a hierarchical manner that is stable over a range of adjacent viewpoints (Marr & Nishihara, 1978; Biederman & Gerhardstein, 1993; Hummel & Stankiewicz, 1996a). Whether a given object representation falls more towards one end of this spectrum or the other may be a product of an observer's experience with the objects in question (Gauthier & Tarr, 1997a; Gauthier, Williams, Tarr, & Tanaka, 1998; Tarr & Bülthoff, 1998). Thus, it is less a question of whether a particular object representation depicts a specific exemplar (e.g., '1965 Mustang') or a larger class (e.g., 'cars') and more how IT-cortex has been fine-tuned by experience to form representations suited for the recognition task most commonly performed by an individual observer (Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999). Our speculation is that regardless of whether one is discriminating between members from within a class (Bülthoff & Edelman, 1992; Humphrey & Khan, 1992; Tarr, 1995) or making more categorical judgments (Biederman, 1987; Hummel & Stankiewicz, 1996b), the critical object representations are comprised of viewpoint-dependent features similar to those that we have begun to explore through computational and psychophysical methods.

View-based representations: Theories of specifically view-based representations in human vision share many overt similarities with computational models of aspect graphs. However, there are also important differences; perhaps the most fundamental being that humans both perceive and represent quantitative as well as qualitative differences among viewpoints. Thus, it is not sur-

prising that view-based representations in humans are sensitive to the statistics of viewpoint (e.g., how often an object is observed in a given viewpoint) and often appear to encode views that do not differ qualitatively (e.g., orientations differing only by a picture-plane rotation).[10] In contrast, an aspect graph representation does not differentiate between such views.

Finally, although the present results do not allow us to directly specify a performance model of biological object recognition, we can speculate on how the geometrical structure captured by aspect graphs might be exploited. One possibility is that insofar as multiple-views representations are sensitive to object geometry, there is a need to delineate boundaries between qualitatively different views. Except for entirely novel objects, a recognizer is likely to encounter two distinct situations: familiar objects in familiar views; and, familiar objects in unfamiliar views. Recognition might then proceed as follows:

1. Configurations of viewpoint-dependent features are located within the input image.
2. In parallel with the location of viewpoint-dependent features, a coarser viewpoint-independent feature-based description is extracted that provides information about category membership and limited information about 3D structure that may facilitate interaction with the object (Biederman, 1987; Ullman, 1989).
3. One or more configurations of features are used as index points, that is, configurations from the input image are compared to configurations encoded in the model base of known objects (Clemens & Jacobs, 1991).
4. The coarse viewpoint-independent description may be used to constrain the search of the model base for viewpoint-dependent feature indexing.

Effectively, what this sequence of events implies is that coarse information about an object, possibly corresponding to basic-level category, will facilitate more specific recognition. In particular, this specificity is achieved by comparing viewpoint-dependent image features to similar features encoded in object memory. Once a correspondence has been established between features in the input and the features in a familiar view, a transformation will be available that may be executed in order to establish a precise match (for instance, see Ullman (1989)). In contrast, for unfamiliar views of objects where the category is extremely certain (e.g., many known exemplars sharing parts with the familiar object), this search will provide a transformation that will yield only an approximate match. That is, the object may still be recognized at both a categorical and at a specific level, but in the latter case, only to the degree that the input image corresponds to familiar views of familiar objects. Finally, for unfamiliar views where the category is less certain, this search will fail to provide a transformation that is likely to yield a good match. Thus, recognition will fail at both levels: coarse-level information being indeterminate and object-specific identification being unattainable.

Each of these three cases also has consequences for whether a new view is encoded: in the first instance, no new view should be instantiated because, while a transformation may have been necessary, there were no qualitative differences between the viewpoint-dependent features of the known view and the perceived view; in the second instance, a new view should be instantiated because of the qualitative differences between the viewpoint-dependent features of the known views *for that object* and the perceived view; and, in the third instance, a new view should be instantiated, but without reference to other pre-existing views, because there is insufficient information to relate this view to known views of objects. Finally, these three situations loosely correspond to psychophysical results: first, mental transformations/normalizations are used to align objects in unfamiliar views with familiar views (Tarr & Pinker, 1989; Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Tarr, 1995); second, views are instantiated according to the visual similarity between an object and other members of a category (Jolicoeur & Milliken, 1989; Moses et al., 1996; Gauthier & Tarr, 1997b; Tarr & Gauthier, 1998); and, third, completely novel views are sometimes treated as entirely new objects (Rock & Di Vita, 1987). Note that nowhere within this hypothetical mechanism is an aspect graph computed per se, rather, the graph structure or multiple-views representation is developed in part according to the qualitative differences, or lack thereof, found among views of objects over experience.

Structural-descriptions: Viewpoint-dependent features may be utilized in mechanisms other than view-based recognition. Consider that many structural-description models for object recognition include recovery mechanisms for extracting 3D volumetric primitives from input shapes (Marr & Nishihara, 1978; Biederman, 1987). Typically, such theories attempt to find invari-

[10] Picture-plane rotations may also produce different qualitative descriptions of objects (for instance, as suggested in Hummel & Biederman, 1992). However, such qualitative changes are *extrinsic* in that they are determined entirely by the perceiver (encompassing both the selected viewpoint and a partitioning of the picture-plane based on statistical regularities in the world, i.e., the object's orientation relative to gravity or functional role). In contrast, the qualitative changes we have examined are *intrinsic* to the object (although a given viewpoint is still extrinsic in being determined by the observer). It is possible that qualitative views, particularly for rotations in the picture-plane, are sometimes defined by purely extrinsic factors (for instance, there is some evidence that arboreal monkeys have cells sensitive to upright monkey faces and other cells sensitive to upside down monkey faces, Perrett, Mistlin, & Chitty, 1987), but that intrinsic factors will predominate or at the least play a significant role in defining qualitative views over rotations in depth.

ants that will support the recovery of an identical 3D description from all viewpoints of an object (disregarding foreshortened or otherwise catastrophic views). Unfortunately, such putative invariants have not shown themselves to be very robust when applied to the range of images likely to be encountered in the 'real world.' Moreover, even less complex contexts may present problems. For example, Biederman (1987) initially suggested that clusters of non-accidental properties could be used to recover 'geons' (a variant of generalized cylinders) independently of orientation; this hypothesis has since been modified to suggest that particular clusters of non-accidental properties are invariant only up to occlusion (Hummel & Biederman, 1992; Biederman & Gerhardstein, 1993). However, recent formal analyses have even rendered this claim suspect-non-accidental properties may not be general enough to support generic recognition (Jacobs, 1992) and empirical results indicate that even single volumes are recognized through viewpoint-dependent processes (Hayward & Tarr, 1997; Tarr et al., 1998).

An alternative approach to recovering volumetric primitives has been proposed by Dickinson, Pentland, and Rosenfeld (Dickinson, Pentland, & Rosenfeld, 1992). Interestingly, their approach combines elements of aspect graph representations with part-based descriptions. They suggest that given a restricted part vocabulary (an attractive proposition for psychological models, see Biederman, 1987), the aspect graphs for the complete set of primitives may be stored. A 3D part description may then be recovered by matching these pre-computed part aspects to viewpoint-dependent features of segmented parts found within input shapes. Thus, an aspect graph representation is used to capture the geometry of parts of objects rather than complete objects. In this manner, view-based mechanisms are used to arrive at a structural-description. This sidesteps some of the problems raised by the aspect graph theory; the number of views per an object and the complexity of the graph structure. Here, the number of aspects is relatively small – primarily because the number of views is independent of object complexity, and in part (*sic*), because of the simplicity of the primitives chosen for the representation.

Two points about this approach should be noted. First, the fact that this approach utilizes view-based features suggests that structural-descriptions are not derived in a serial fashion prior to image-based matching. Rather, it may be that both elements of the representation are computed concurrently with a degree of interaction; thus, partial information about category may constrain more specific indexing, while partial information about image-based matches may constrain the recovery of more complex descriptors of the object (e.g., configurations of features or parts). Second, this method is quite similar to earlier structural-description approaches in that it relies on a restricted set of primitives (in order to limit the number of aspects) and on using viewpoint-dependent features for recovering such primitives. In particular, because the features currently hypothesized for geon recovery may not prove robust in the more general case, it may be worthwhile to consider alternative classes of image features; significantly, because of both their generality and empirically validated salience, we suggest that the features used in aspect graphs are good candidates.

5.2. Other viewpoint-dependent factors

While the results of our experiments indicate that observers are sensitive to some of the viewpoint-dependent features used in computing aspect graphs, there are both limitations to this claim and additional viewpoint-dependent factors almost certainly enter into the definition of aspects. First, the difference in performance between the line drawings and shaded images raises a number of issues about the saliency of certain configurations of features under 'normal' viewing conditions (a point also relevant to the recovery of geons from 'non-accidental' properties). Most notably, observers were not as adept at distinguishing orientation changes across the swallowtail transitions (the events at 45° and 135°) of the shaded version of the torus as compared to the line drawing version (Fig. 6). Difficulty in perceiving this change for shaded drawings may be related to the accuracy of locating (or even detecting) cusps in images. In particular, the response of an edge detection filter diminishes in the neighborhood of a cusp: because an edge terminates at a cusp and a receptive field has non-zero area, the edge will cross only one-half of the receptive field when it is centered at the cusp point. In the case of a swallowtail transition, a cusp and a t-junction are introduced; near the accidental viewpoint, the cusp will be quite close to the junction and, consequently, will be very difficult to detect given the limitations of human edge detectors. In light of this fact, it is not at all surprising that observers did not consistently discriminate between the two views in the neighborhood of the swallowtail.

Similarly, in the case of the bell, observers were poorer at discriminating between the views across the visual event occurring at 151.2° in the shaded image versus the line drawing. Compare Fig. 9(a) and (b). In this visual event, a cusp and a t-junction merge into a smooth contour; illustrated in the first transition of Fig. 11. Again, difficulty in discerning cusps may account for the relatively poorer performance.

Here, we see limitations of the assumptions used to define the aspect graph as described earlier. Recall that the second of the three issues that needed to be addressed was 'What constitutes a qualitative description of a view?' In this discussion, as in nearly all work

concerning aspect graphs, views were defined by the features to be found in perfect line drawings. However, as with many competence models, this does not consider the limitations of the actual perceptual system, e.g., performance, in this instance the edge detection process, and in particular, the fact that receptive fields have non-zero area. To account for this effect, a plausible performance theory encompassing aspect graphs will require significant modifications; beyond even those proposed in recent articles that include degenerate views (Kender & Freudenstein, 1987) and finite resolution aspect graphs (Eggert et al., 1993; Shimshoni & Ponce, 1997).

A second limitation of the current framework concerns the purely qualitative definition of a view. As such, no distinction is made between viewpoints that are within a single region, but result in significant quantitative changes in the image of the object. As discussed, there is some evidence within our data that observers are sensitive to certain quantitative changes and may utilize them in delineating among viewpoints. In particular, observers may be sensitive to the shape of the silhouette and the total area it encompasses (Hayward & Tarr, 1997). Supporting this interpretation, there are several studies that have also explored the factors that determine aspects of objects in human perception and come to similar conclusions.

First, Perrett and Harries (1988) recorded how participants distribute their time while freely inspecting objects rotated around a vertical axis. Using both tetrahedra and smooth objects (potatoes), they found that participants were more likely to spend greater amounts of time studying both the maximum and minimum horizontal width of the silhouette (the viewpoints where a principle axis of the object was either parallel to the image plane or maximally foreshortened). This result is consistent with our finding that within certain qualitative regions, our observers' ability to discriminate between viewpoints increases with changes in the silhouette of the object. However, Perrett and Harries' experiments also revealed large individual differences among participants in terms of which viewpoints were preferred; a finding that suggests that simple exploration experiments may not adequately capture the underlying mechanisms used to define and instantiate view-based representations. Indeed, an exploration paradigm is subject to both individual preferences and specific experiences with different views of objects. Different views of the same object may be selected as the 'best' depending on subtle differences in how the task is described (Blanz et al., 1999). A second study by Harries, Perrett, and Lavender (1991) supports this conjecture. Using 3D heads as stimuli, they found that differential inspection times across viewpoints did not predict differences in recognition performance or memory. On closer inspection, these findings may still be compatible with a multiple-views explanation; it is possible that participants learn prototypical views of heads (as reflected in the distribution of inspection times and through prior experience with heads as a class), but that such prototypical views essentially 'cover' the entire view space such that unencoded intermediate views activate responses in two more adjacent prototypical views (as reflected in equivalent or even better recognition performance at intermediate views; Tarr & Pinker, 1989; Bülthoff & Edelman, 1992). Interestingly, this interpretation is consistent with computational approaches in which unfamiliar views are recognized by a linear combination (Ullman & Basri, 1991) or of an interpolation (Poggio & Edelman, 1990) between two nearby views.

Second, Langdon, Mayhew, and Frisby (1991) had participants rate the difference between an object appearing in a reference view and in views generated by varying degrees of rotation around a variety of different axes. A simple measure of feature difference was used to assess whether object geometry played a role in determining ratings. This measure was compiled by comparing the cumulative number of faces, edges, and vertices in the reference view to the features in the rated view. Note that while this measure does correlate somewhat with qualitative changes in the view space, it is imperfect in that some qualitative differences will be missed. Results from this study were consistent with the hypothesis that observers are sensitive to both quantitative and qualitative changes in view. On the one hand, difference ratings generally increased with greater angular separation between orientations, suggesting that quantitative changes help determine the perceived viewpoint of an object; on the other hand, discontinuities in the ratings occurred at viewpoints that often corresponded to large transitions in the measured feature differences, suggesting that qualitative changes also played a part in determining the perceived viewpoint. Thus, there is at least limited evidence for somewhat richer view-based representations in humans than as specified by aspect graph theory.

5.3. Conclusions

To summarize the ideas we have presented here, we wish to emphasize the following points.

View-based representations seem to underlie many of the mechanisms used by humans to visually recognize objects. In particular, such representations are particularly engaged when discriminating between visually-similar objects (Tarr & Pinker, 1989, 1990; Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Humphrey & Khan, 1992; Tarr, 1995).

Factors that influence the acquisition of view-based representations include familiarity with a given view of an object, similarity of a given view of an object to

known views of visually-similar objects, and the way in which the geometry of an object varies with viewpoint (Gauthier & Tarr, 1997b; Tarr & Gauthier, 1998). These factors reflect the fact that humans are most likely sensitive to both quantitative and qualitative differences in views.

The way in which image geometry varies with viewpoint may be captured by an aspect graph representation providing a complete decomposition of an object into its aspects. Such views are delineated by a small class of qualitative changes that occur in configurations of image features for changes in viewpoint (Koenderink & Van Doorn, 1979; Koenderink, 1990; Kriegman & Ponce, 1990).

Human perceivers are sensitive to several of the viewpoint-dependent features used to construct aspect graphs. While it is unlikely that observers 'construct' aspect graphs, it is possible that such features provide the underpinnings of mechanisms that determine what constitutes a unique view of an object in view-based representations. Furthermore, it is also possible that the same features may provide the basis for other aspects of object representation; for instance, the robust recovery of parts.

Acknowledgements

Support for MJT was provided by the Air Force Office of Scientific Research, contract number F49620-91-J-0169. DJK was supported in part by a NSF Young Investigator Award, IRI-9257990. Support to both authors was provided by the Office of Naval Research contract number N00014-93-1-0305. We would like to thank Jean Ponce of the University of Illinois with whom the methods for computing aspect graphs from object models were developed. We thank Steve Messe, KaRin Turner, Scott Yu for their assistance in running the psychophysical studies, Laurie M. Heller, Robert Bjork, Steve Palmer, and Irving Biederman for their helpful comments, and Scott Yu and Mohsin Malik for the illustrations of how contours on surfaces interact.

References

Arnol'd, V. (1984). *Catastrophe theory*. Heidelberg: Springer.

Bartram, D. J. (1974). The role of visual and semantic codes in object naming. *Cognitive Psychology*, *6*, 325–356.

Bartram, D. J. (1976). Levels of coding in picture–picture comparison tasks. *Memory & Cognition*, *4*, 593–602.

Beymer, D., & Poggio, T. (1996). Image representations for visual learning. *Science*, *272*, 1905–1909.

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

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(6), 1162–1182.

Biederman, I., & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1506–1514.

Blanz, V., Tarr, M. J., & Bülthoff, H. H. (1999). What object attributes determine canonical views? *Perception*, *28*, 575–600.

Bowyer, K., & Dyer, C. R. (1991). Aspect graphs: an introduction and survey of recent results. *International Journal of Imaging Systems and Technology*, *2*, 315–328.

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

Chen, S., & Freeman, H. (1991). On the characteristic views of quadric-surfaced solids. In IEEE workshop on directions in automated CAD-Based vision (pp. 34–43). Hawaii: Maui.

Clemens, D. T., & Jacobs, D. W. (1991). Space and time bounds on indexing 3-D models from 2-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(10), 1007–1017.

Corballis, M. C. (1988). Recognition of disoriented shapes. *Psychological Review*, *95*(12), 115–123.

Dickinson, S. J., Pentland, A. P., & Rosenfeld, A. (1992). From volumes to views: an approach to 3-D object recognition. *Computer Vision, Graphics, and Image Processing: Image Understanding*, *55*(2), 130–154.

Edelman, S. (1995). Representation, similarity, and the chorus of prototypes. *Minds and Machines*, *5*(1), 45–68.

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

Edelman, S., & Weinshall, D. (1991). A self-organizing multiple-view representation of 3D objects. *Biological Cybernetics*, *64*, 209–219.

Eggert, D., & Bowyer, K. (1993). Computing the perspective projection aspect graph of solids of revolution. *IEEE Transactions of Pattern Analysis and Mechanical Intelligence*, *15*(2), 109–128.

Eggert, D., Bowyer, K., Dyer, C., Christensen, H., & Goldgof, D. (1993). The scale space aspect graph. *IEEE Transactions of Pattern Analysis and Mechanical Intelligence*, *15*(11), 1114–1131.

Ellis, R., & Allport, D. A. (1986). Multiple levels of representation for visual objects: a behavioural study. In A. G. Cohn, & J. R. Thomas, *Artificial intelligence and its applications* (pp. 245–247). New York: Wiley.

Farah, M. J. (1992). Is an object an object an object? Cognitive and neuropsychological investigations of domain-specificity in visual object recognition. *Current Directions in Psychological Science*, *1*(5), 164–169.

Freeman, H., & Chakravarty, I. (1980). The use of characteristic views in the recognition of three-dimensional objects. In E. S. Gelsema, & L. N. Kanal, *Pattern recognition in practice* (pp. 277–288). New York: North Holland.

Gauthier, I., & Tarr, M. J. (1997a). Becoming a 'Greeble' expert: exploring the face recognition mechanism. *Vision Research*, *37*(12), 1673–1682.

Gauthier, I., & Tarr, M. J. (1997b). Orientation priming of novel shapes in the context of viewpoint-dependent recognition. *Perception*, *26*, 51–73.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2*(6), 568–573.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training 'Greeble' experts: a framework for studying expert object recognition processes. *Vision Research*, *38*, 2401–2428.

Gigus, Z., & Malik, J. (1990). Computing the aspect graph for line drawings of polyhedral objects. *IEEE Transactions in Pattern Analysis and Mechanical Intelligence*, *12*(2), 113–122.

Harries, M. H., Perrett, D. I., & Lavender, A. (1991). Preferential inspection of views of 3-D model heads. *Perception*, *20*, 669–680.

Hayward, W. G., & Tarr, M. J. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(5), 1511–1521.

Hebert, M., & Kanade, T. (1985). The 3D profile method for object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 458–464). San Francisco, CA.

Hummel, J. E., & Biederman, L. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480–517.

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

Hummel, J. E., & Stankiewicz, B. J. (1996b). An architecture for rapid, hierarchical structural description. In T. Inui, & J. McClelland, *Attention and performance XVI* (pp. 93–121). Cambridge, MA: MIT Press.

Humphrey, G. K., & Khan, S. C. (1992). Recognizing novel views of three-dimensional objects. *Canadian Journal of Psychology*, *46*, 170–190.

Ikeuchi, K., & Kanade, T. (1988). Automatic generation of object recognition programs. *Proceedings of the IEEE*, *76*(8), 1016–1035.

Ikeuchi, K., & Kanade, T. (1989). Modeling sensors: toward automatic generation of object recognition program. *Computer Vision, Graphics, and Image Processing*, *48*(1), 50–79.

Jacobs, D. W. (1992). Space efficient 3D model indexing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 439–444).

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory and Cognition*, *13*, 289–303.

Jolicoeur, P. (1990). Identification of disoriented objects: a dual-systems theory. *Mind and Language*, *5*(4), 387–410.

Jolicoeur, P., & Milliken, B. (1989). Identification of disoriented objects: effects of context of prior presentation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 200–210.

Kender, J. R., & Freudenstein, D. G. (1987). What is a 'degenerate' view? In *Proceedings of the DARPA image understanding workshop*, *Vol. 11*, (pp. 589–598). Morgan Kaufmann, Las Altos, CA.

Kergosien, Y. L. (1981). La famille des projections orthogonales d'une surface et ses singularités. *C.R. Academy of Science Paris*, *292*, 929–932.

Koenderink, J. J. (1984). What does the occluding contour tell us about solid shape? *Perception*, *13*(3), 321–330.

Koenderink, J. J. (1990). *Solid shape*. Cambridge, MA: MIT Press.

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

Koenderink, J. J., & Van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics*, *32*, 211–216.

Kriegman, D. J., & Ponce, J. (1990). Computing exact aspect graphs of curved objects: solids of revolution. *International Journal of Computer Vision*, *5*(2), 119–135.

Lando, M., & Edelman, S. (1995). Receptive field spaces and class-based generalization from a single view in face recognition. *Network*, *6*, 551–576.

Langdon, P. M., Mayhew, J. E. W., & Frisby, J. P. (1991). In search of 'characteristic view' 3D object representations in human vision using ratings of perceived differences between views. In J. E. W. Mayhew, & J. P. Frisby, *3D model recognition from stereoscopic cues* (pp. 245–248). Cambridge, MA: MIT Press.

Lawson, R., Humphreys, G. W., & Watson, D. G. (1994). Object recognition under sequential viewing conditions: evidence for viewpoint-specific recognition procedures. *Perception*, *23*(5), 595–614.

Layman, S., & Greene, E. (1988). The effect of stroke on object recognition. *Brain and Cognition*, *7*, 87–114.

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

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

Malik, J. (1987). Interpreting line drawings of curved objects. *International Journal of Computer Vision*, *1*(1), 73–103.

Marr, D. (1982). *Vision*: *a computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

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

Moses, Y., Ullman, S., & Edelman, S. (1996). Generalization to novel images in upright and inverted faces. *Perception*, *25*, 443–462.

Nalwa, V. (1988). Line-drawing interpretation: a mathematical framework. *International Journal of Computer Vision*, *2*, 103–124.

Norman, J. F., & Todd, J. T. (1994). Perception of rigid motion in depth from the optical deformations of shadows and occlusion boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 343–356.

Norman, J. F., Todd, J. T., & Phillips, F. (1995). The perception of surface orientation from multiple sources of optical information. *Perception & Psychophysics*, *57*(5), 629–636.

Pae, S., & Ponce, J. (1999). Toward a scale-space aspect graph: solids of revolution. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. II: 196–201). Ft. Collins, CO.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long, & A. Baddeley, *Attention and performance IX*. Hillsdale, NJ: Lawrence Erlbaum.

Perrett, D. I., & Harries, M. H. (1988). Characteristic views and the visual inspection of simple faceted and smooth objects: 'tetrahedra and potatoes'. *Perception*, *17*, 703–720.

Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., Chitty, A. J., Hietanen, J. K., & Ortega, J. E. (1989). Frameworks of analysis for the neural representations of animate objects and actions. *Journal of Experimental Biology*, *146*, 87–113.

Perrett, D. I., Harries, M. H., & Looker, S. (1992). Use of preferential inspection to define the viewing sphere and characteristic views of an arbitrary machined tool part. *Perception*, *21*, 497–515.

Perrett, D. I., Mistlin, A. J., & Chitty, A. J. (1987). Visual neurones responsive to faces. *Trends in Neuroscience*, *10*(96), 358–364.

Perrett, D. I., Oram, M. W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition*, *67*(1, 2), 111–145.

Perrett, D. I., Oram, M. W., Harries, M. H., Bevan, R., Hietanen, J. K., Benson, P. J., & Thomas, S. (1991). Viewer-centred and object-centred coding of heads in the macaques temporal cortex. *Experimental Brain Research*, *86*, 159–173.

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

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

Petitjean, S., Ponce, J., & Kriegman, D. J. (1992). Computing exact aspect graphs of curved objects: Algebraic surfaces. *International Journal of Computer Vision*, *9*(3), 231–255.

Phillips, F., Todd, J. T., Koenderink, J. J., & Kappers, A. M. L. (1997). Perceptual localization of surface position. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(5), 1481–1492.

Pirsig, R. M. (1974). *Zen and the art of motorcycle maintenance*. Toronto, Canada: Bantam Books.

Plantinga, H., & Dyer, C. (1990). Visibility, occlusion, and the aspect graph. *International Journal of Computer Vision*, *5*(2), 137–160.

Platonova, O. (1984). Singularities of projections of smooth surfaces. *Russian Mathematical Surveys*, *39*, 177–178.

Plaut, D. C., & Farah, M. J. (1990). Visual object representation: interpreting neurophysiological data within a computational framework. *Journal of Cognitive Neuroscience*, *2*(4), 320–343.

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

Poggio, T., & Vetter, T. (1992). Recognition and structure from one 2D model view: observations on prototypes, object classes and symmetries (Tech. Rep. No. 1347), Massachusetts Institute of Technology.

Ratcliff, G., & Newcombe, F. A. (1982). Object recognition: some deductions from the clinical evidence. In A. W. Ellis, *Normality and pathology in cognitive functions* (pp. 147–171). New York: Academic Press.

Rieger, J. (1987). On the classification of views of piecewise smooth objects. *Image and Vision Computing*, *5*, 91–97.

Rieger, J. (1990). The geometry of view space of opaque objects bounded by smooth surfaces. *Artificial Intelligence*, *44*, 1–40.

Rieger, J. (1992). Global bifurcations sets and stable projections of non-singular algebraic surfaces. *International Journal of Computer Vision*, *7*(3), 171–194.

Riesenhuber, M., & Dayan, P. (1997). Neural models for part-whole hierarchies. In *Advances in neural information processing systems*, vol. 9 (pp. 17–23). Cambridge, MA: MIT Press.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Rock, I. (1973). In: 'Orientation and form' New York: Academic Press.

Rock, I. (1974). The perception of disoriented figures. *Scientific American*, *230*, 78–86.

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

Sanocki, T., Bowyer, K. W., Heath, M. D., & Sarkar, S. (1998). Are edges sufficient for object recognition? *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 340–349.

Seibert, M., & Waxman, A. M. (1990). Learning aspect graph representations from view sequences. In D. S. Touretzky, *Advances in neural information processing systems* 2 (pp. 258–265). Los Altos, CA: Morgan Kaufmann.

Seibert, M., & Waxman, A. M. (1992). Learning and recognizing 3D objects from multiple views in a neural system. In H. Wechsler, *Neural networks for perception* (pp. 428–444). New York: Academic Press.

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.

Shimshoni, I., & Ponce, J. (1997). Finite-resolution aspect graphs of polyhedral objects. *IEEE Transaction Pattern Analysis and Mechanics Intelligence*, *19*(4), 315–327.

Sripradisvarakul, T., & Jain, R. (1989). Generating aspect graphs of curved objects. In *IEEE workshop on interpretation of 3D scenes* (pp. 109–115). Austin, TX.

Stewman, J., & Bowyer, K. (1987). Aspect graphs for planar-face convex objects. In: IEEE workshop on computer vision (pp. 123–130).

Tarr, M. J. (1995). Rotating objects to recognize them: a case study of the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review*, *2*(1), 55–82.

Tarr, M. J., & Bülthoff, H. H. (1995). Is human object recognition better described by geon-structural-descriptions or by multiple-views? *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1494–1505.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey, and machine. *Cognition*, *67*(1–2), 1–20.

Tarr, M. J., & Gauthier, I. (1998). Do viewpoint-dependent mechanisms generalize across members of a class? *Cognition*, *67*, 71–108.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*(28), 233–282.

Tarr, M. J., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science*, *1*(42), 253–256.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint-dependent. *Nature Neuroscience*, *1*(4), 275–277.

Thom, R. (1972). *Structural stability and morphogenesis*. New York: Benjamin.

Todd, J. T., Chen, L., & Norman, J. F. (1998). On the relative salience of euclidean, affine, and topological structure for 3-d form discrimination. *Perception*, *27*(3), 273–282.

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 Transactions on Pattern Analysis and Machine Intelligence*, *13*(10), 992–1006.

Van Effelterre, T. (1994). Aspect graphs for visual recognition of three-dimensional objects. *Perception*, *23*, 563–582.

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

Warman, M., Baugher, S., & Gualtieri, J. (1986). The visual potential for one convex polygon (Center for Automation Research No. CAR-TR-212), University of Maryland.

Warrington, E. K., & Taylor, A. M. (1973). The contribution of right parietal lobe to object recognition. *Cortex*, *9*, 152–164.

Watts, N. (1987). Calculating the principal views of a polyhedron (Tech. Rep. No. CS Tech. Report 234), Rochester University.

Weinshall, D., Edelman, S., & Bülthoff, H. H. (1990). A self-organizing multiple-view representation of 3D objects. In D. S. Touretzky, *Advances in neural information processing systems* 2 (pp. 274–281). Los Altos, CA: Morgan Kaufmann.

Whitney, H. (1955). On singularities of mappings of Euclidean spaces. I. Mappings of the plane into the plane. *Annals of Mathematics*, *62*(3), 374–410.