ELSEVIER

Cognition 67 (1998) 73–110

COGNITION

Do viewpoint-dependent mechanisms generalize across members of a class?

Michael J. Tarr[a,]*, Isabel Gauthier[b]

[a]*Department of Cognitive and Linguistic Sciences, Brown University, P.O. Box 1978, Providence, RI 02912, USA*
[b]*Department of Psychology, Yale University, New Haven, CT 06520, USA*

Abstract

Evidence for viewpoint-specific image-based object representations have been collected almost entirely using exemplar-specific recognition tasks. Recent results, however, implicate image-based processes in more categorical tasks, for instance when objects contain qualitatively different 3D parts. Although such discriminations approximate class-level recognition, they do not establish whether image-based representations can support generalization across members of an object class. This issue is critical to any theory of recognition, in that one hallmark of human visual competence is the ability to recognize unfamiliar instances of a familiar class. The present study addresses this question by testing whether viewpoint-specific representations for some members of a class facilitate the recognition of other members of that class. Experiment 1 demonstrates that familiarity with several members of a class of novel 3D objects generalizes in a viewpoint-dependent manner to cohort objects from the same class. Experiment 2 demonstrates that this generalization is based on the degree of familiarity and the degree of geometrical distinctiveness for particular viewpoints. Experiment 3 demonstrates that this generalization is restricted to visually-similar objects rather than all objects learned in a given context. These results support the hypothesis that image-based representations are viewpoint dependent, but that these representations generalize across members of perceptually-defined classes. More generally, these results provide evidence for a new approach to image-based recognition in which object classes are represented as clusters of visually-similar viewpoint-specific representations. © 1998 Elsevier Science B.V. All rights reserved

Keywords: Class generalization; Image-based recognition; Viewpoint-specific representation

* Corresponding author. Tel.: +1 401 8631148; fax: +1 401 8632255; e-mail: Michael_Tarr@brown.edu

PII S0010-0277(98)00023-7

1. Introduction

A significant body of work on human object recognition has been concerned with the question of how observers recognize objects from unfamiliar viewpoints (Rock, 1973). Recent results suggest that there is no definitive answer to this question, rather there is a continuum ranging from extreme viewpoint dependence to almost complete viewpoint invariance. There is, however, a general principle underlying this continuum: The *degree* of viewpoint dependence is largely determined by the between-item similarity of objects that must be discriminated, with more homogeneity between objects leading to greater viewpoint dependence and less homogeneity leading to less viewpoint dependence (Tarr and Bülthoff, 1995; Schyns, 1998). This claim appears to hold across a wide range of stimuli and tasks, including studies using alphanumeric characters (Corballis et al., 1978), common objects (Bartram, 1974; Jolicoeur, 1985; Lawson et al., 1994), novel 2D (Tarr and Pinker, 1989) and 3D objects (Bülthoff and Edelman, 1992; Humphrey and Khan, 1992; Biederman and Gerhardstein, 1993; Tarr, 1995; Hayward and Tarr, 1997; Tarr et al., 1997), or faces (Yin, 1969; Troje and Bülthoff, 1996; Hill et al., 1997).

Given such mixed results, different theorists have drawn quite different conclusions regarding the mechanisms used for visual recognition (Biederman and Gerhardstein, 1995; Tarr and Bülthoff, 1995). On the one hand, relatively smaller effects of viewpoint (e.g. Corballis et al., 1978; Biederman and Gerhardstein, 1993) have typically been interpreted as evidence for a structural-description system in which objects are represented as assemblies of 3D parts that are stable over large changes in viewpoint (Marr and Nishihara, 1978; Biederman, 1987). On the other hand, relatively large effects of viewpoint (e.g. Bülthoff and Edelman, 1992; Tarr, 1995) have been interpreted as evidence for an image-based or view-based system in which objects are represented as sets of metrically-specific features that are unstable over changes in viewpoint (Poggio and Edelman, 1990; Bülthoff et al., 1995). While both types of theories offer parsimonious accounts for some subset of the data, there remains the larger question of the domain covered by each. One common reconciliation has been to assume only limited domains for each type of mechanism: basic- or entry-level recognition of object category being handled by a qualitative part-based system, while subordinate-level recognition of specific exemplars being handled by a quantitative image-based system (Jolicoeur, 1990).

By some views, this solution is less than satisfactory and inconsistent with at least some of the extant data. Consequently, theorists have begun to hypothesize that recognition is almost entirely part-based (Biederman and Gerhardstein, 1993, 1995) or almost entirely image-based (Edelman, 1995; Tarr and Bülthoff, 1995). What is still unclear is how each approach can be extended to accommodate recognition tasks that were not part of the original domain of explanation. Indeed, proponents of the part-based approach have explicitly criticized the generality of image-based theories on this basis, suggesting that such mechanisms are incapable of generalizing across unfamiliar instances of familiar classes, that is, entry-level recognition (Biederman and Gerhardstein, 1993, 1995). In contrast, proponents of the image-based approach have proposed schemes in which viewpoint-specific representations do generalize to new members of familiar classes (Poggio and Brunelli, 1992; Lando and Edelman, 1995; Vetter et al., 1995; Beymer and Poggio, 1996; Moses et al., 1996; Gauthier and Tarr, 1997b). Empirical evidence on this issue is, however, somewhat thin. Thus, the goal of this paper is to investigate the nature of image-based class generalization, asking: (1) Does such generalization occur? (2) What factors mediate generalization? (3) Is this generalization based on the same visual similarity that helps to define a visual class?

1.1. Evidence for generalization

Tests of viewpoint dependence in recognition have focused on whether subjects learn specific objects in specific views. In a typical experiment there is an initial viewpoint dependency that diminishes with extensive practice to near-equivalent performance at familiar views (Tarr and Pinker, 1989; Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992; Humphrey and Khan, 1992; Tarr, 1995). This near-invariance may be interpreted as evidence for either viewpoint-invariant or multiple viewpoint-specific object representations. Critically, Tarr and Pinker (1989) and since then Bülthoff and Edelman (1992) and Tarr (1995) found that performance for *unfamiliar* views remained viewpoint dependent and, moreover, was related to the distance from the nearest familiar view.

Interestingly, Jolicoeur and Milliken (1989) obtained diminished effects of viewpoint at unfamiliar viewpoints, reminiscent of those found after extensive practice, *without* the benefit of subjects actually viewing the specifically tested objects in the test viewpoints. Their subjects, however, did view *other* objects at the tested viewpoints, suggesting that viewpoint invariance may be produced by the context of the prior presentation of different objects – cohorts – at the same test viewpoints subsequently used to assess viewpoint invariance (Jolicoeur and Milliken, 1989). As with diminished effects of viewpoint due to practice, diminished effects due to context may be accounted for by either viewpoint-invariant or multiple viewpoint-specific representations. Largely because subjects never observed a given object (only its cohorts) at the viewpoints in question, Jolicoeur and Milliken interpreted their results as evidence for viewpoint-invariant mechanisms, as have subsequent transfer experiments (see Murray et al., 1993). An alternative interpretation of this result is that viewpoint-specific image-based representations formed for the objects actually seen at a given viewpoint may generalize to visually-similar objects seen only later at the same view, that is, image-based class generalization.

1.2. Image-based class generalization

A problem with image-based theories is that views are typically assumed to be specific to exact images features and attributes. For example, in the influential neural-network model by Poggio and Edelman (1990) (see also Weinshall et al., 1990; Edelman and Weinshall, 1991) objects were coded by the precise (x, y) coordinates of their vertices. Such a coding is both impractical and at odds with our intuitions regarding object recognition. Take for example, the typical real-world situation in which new exemplars of a familiar category are seen for the first time, e.g. a new model of car. Our intuitions tell us that our knowledge about cars we have seen in the past facilitates our recognition of this new car. In Poggio and Edelman's model, however, such generalization could not occur – the representation of one car would be specific to a given set of coordinates and, thus, would not match a new visually-similar car (although the Edelman and Weinshall model may be able to handle this case due to blurring of the input). In contrast, it seems that our knowledge about an entire category facilitates the recognition process. Thus, it may be possible to recognize a new exemplar of a known category from a novel view based on the knowledge of the class. Similar intuitions have led many proponents of the image-based approach to develop computational models for generalizing between members of a homogeneous class using viewpoint-specific representations (Poggio and Brunelli, 1992; Lando and Edelman, 1995; Vetter et al., 1995; Beymer and Poggio, 1996; Moses et al., 1996).

There is already considerable empirical evidence for image-based recognition mechanisms. What is unknown is whether the obvious strength of this approach, the coding of metric specificity that can support exemplar-specific recognition, can be retained while extending it to handle class-level recognition. One possible model for doing this, advocated by Gauthier and Tarr (1997b) (see also Edelman, 1995), involves pooling activation across a number of visually-similar image-based representations (for neural evidence for models of this sort, see Perrett et al., 1998). The idea is that classes may be represented as clusters of exemplar-specific and, crucially, viewpoint-specific image-based views that can support generalization from one exemplar to another. There are at least two sources of empirical support for this model. First, Moses et al. (1996) found that observers were good at generalizing from a single view of an unfamiliar upright face to new views of the same upright face, but were poor at generalizing from single views of inverted faces to new inverted views. This finding suggests that humans appear to have a class-general, viewpoint-specific, i.e. upright only, representation for faces. Second, Gauthier and Tarr (1997b) found that viewpoint-specific representations of visually-similar novel 2D shapes interacted to facilitate recognition. Specifically, we found evidence for *orientation priming* – better recognition of a shape at a particular orientation based on the prior presentation of other visually-similar shapes at the same orientation. In other words, given shapes from a homogeneous[1] class, e.g. S1, S2 and S3, learned at 0°, recognition of shapes S1 and S2 at 120° reduced the effect of viewpoint for the subsequent recognition of shape S3 at 120°. These results were limited, however, by the fact that the shapes were rotated only in the picture-plane and that orientation priming occurred quite early in testing – most likely before subjects could have acquired new object-specific representations at 120°. In research presented here we wished to explore 3D class generalization more directly, that is, in conditions where we were sure that subjects had learned object O1 at viewpoints α and θ and object O2 only at viewpoint α. We hypothesize that once a view has been learned for O1 at

[1] Throughout the paper, we use an informal definition of homogeneous and visually-similar. For purposes of the experiments presented here, we need only assume that the perceptual information that subjects rely on for object recognition overlaps across objects defined as similar.

θ, recognition performance for the visually-similar O2 will be enhanced at θ via image-based generalization.

In all three experiments the logic for testing this prediction is similar to that used for assessing whether diminished effects of viewpoint with practice are due to viewpoint-invariant or viewpoint-dependent mechanisms (Tarr and Pinker, 1989). The crucial difference here is that familiar test objects are now presented at viewpoints at which only their visually-similar cohorts have appeared previously (*cohort views*), as well as at unfamiliar viewpoints where neither the test objects nor their cohorts have appeared previously (*novel views*). Two outcomes are possible for performance at these test viewpoints.

(1) Response times are equivalent (and fast) at cohort and novel views. Such a result would support a viewpoint-invariant interpretation, suggesting that the appearance of only *some* members of a class at several viewpoints prompts subjects to acquire more general viewpoint-invariant representations. This is in contrast to earlier studies where *all* objects appeared at the same subset of viewpoints, apparently prompting subjects to acquire viewpoint-specific representations at each highly familiar viewpoint (Tarr and Pinker, 1989; Tarr, 1995).

(2) Response times for cohort views are fast, while response times for novel views are systematically related to the distance from the nearest familiar *or* cohort view. Such a result would support a viewpoint-dependent interpretation, suggesting that the appearance of only (*some*) members of a class at several viewpoints prompts subjects to acquire viewpoint-specific, but class-general, representations. Thus, as in earlier studies where highly familiar viewpoint-specific representations served as direct matches or targets for normalization processes, cohort views may serve similarly, but for objects never actually seen at those viewpoints.

2. Experiment 1

Experiment 1 examines whether learning viewpoint-specific information about novel objects generalizes to other visually-similar objects. Based on the earlier results of Jolicoeur and Milliken (1989) we know that naming familiar objects at a given orientation facilitates the naming of other familiar objects at the same orientation (see also Murray et al., 1993). However, because they failed to probe orientations that were unfamiliar for all of the objects, it is unclear whether the facilitation Jolicoeur and Milliken obtained is mediated by viewpoint-invariant or viewpoint-dependent mechanisms. We address this question by measuring recognition performance at familiar, cohort, and novel views, as well as extending Jolicoeur and Milliken's findings to novel 3D objects and rotations in depth.

2.1. *Method*

2.1.1. Subjects

Twelve subjects from the MIT community participated in the experiment for pay. All reported normal or corrected to normal vision.

2.1.2. Materials

The complete stimulus set of seven objects is illustrated in Fig. 1 at their arbitrarily designated canonical viewpoint of 10° around each of the three principle axes (Fig. 2); (see Tarr (1995) for details on how the objects were generated). As illustrated in Fig. 2 objects were rotated around either the *X*, *Y*, or *Z* axis with the other two axes held constant at 10° (the order of rotations was always *X*, *Y*, and *Z*). A complete set of 34 viewpoints (including the canonical viewpoint) was generated by rotating each object through eleven 30° intervals (40°, 70°, ... , 340°) around the *X*, *Y*, or *Z* axis.

Rotations were centered around the geometric midpoint of the object as defined by the furthest reaches of its arms. Stimuli were displayed centered on a color EGA monitor at a resolution of 512 × 512 pixels within a circle approximately 13 cm in diameter (19.4° of visual angle). The surfaces of the stimuli were colored a uniform light blue and the edges of the faces of each cube were colored red with hidden lines removed (for further details, see Tarr, 1995).

2.1.3. Design and procedure

In the *Training* phase subjects learned the names and 3D structure for a subset of four target objects at the canonical training viewpoint (Fig. 1). Subjects learned the objects by copying them and then building them from memory using a construction toy that allowed them to attach single units to the main axis fixed at the training

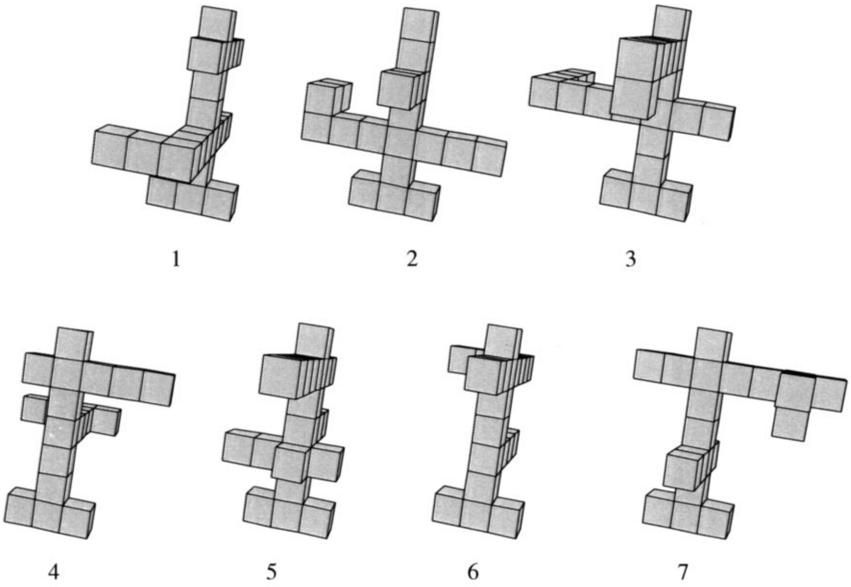

Fig. 1. The set of novel 3D objects used as stimuli in Experiments 1 and 2. The objects are shown at their near-upright training viewpoint (10°, 10°, 10°). Note that the set of objects form a somewhat homogeneous visual class in that they share common components (cubes) and a clearly marked bottom 'foot' and major vertical axis.

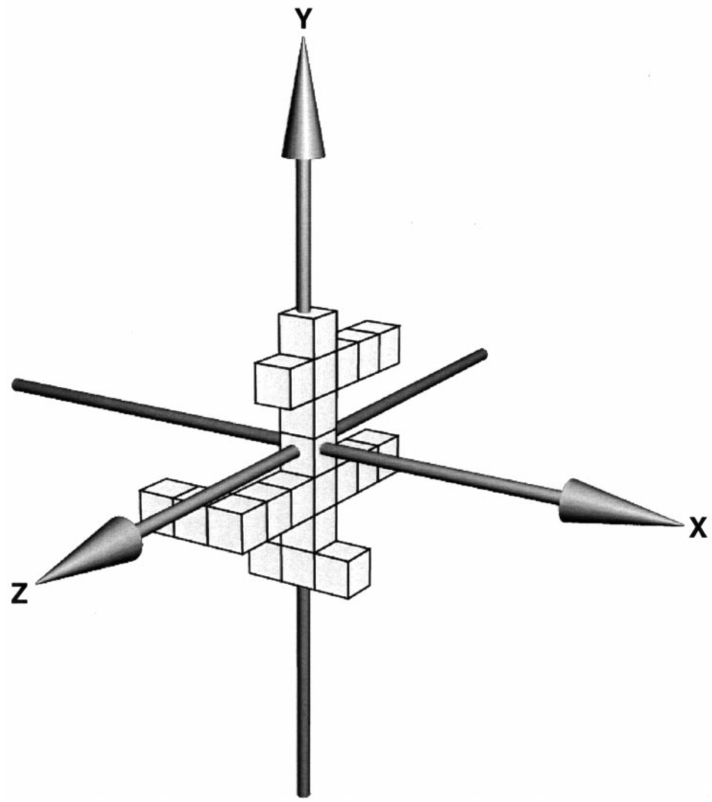

Fig. 2. Axes of rotation used to generate changes in viewpoint. For Experiments 1 and 2 each object was rotated around either the *X*, *Y* or *Z* axis with the other two axes held constant at 10°. For Experiment 3 each object was rotated only around the *Y* axis with the other two axes held constant at 0°. For all experiments rotations were centered around the geometric midpoint of the bounding box of a given object (as defined by the furthest reaches of the arms).

viewpoint (from their perspective). When subjects failed to correctly build an object they were given feedback as to where they made errors. Subjects were trained until they could twice successfully build all four of the objects from memory.

In the *Practice* phase subjects were shown the objects on the monitor at one of several select viewpoints and practiced naming each target object by pressing response keys labeled with their corresponding names. Other members of the complete stimulus set were presented as distractors and subjects responded by pressing a foot pedal. A subset of the objects, including two named targets and three unnamed distractors appeared at ten different viewpoints (the Rotated set), while the remaining two named targets appeared only at the training viewpoint (the Unrotated set). The viewpoints in which the Rotated set appeared, but the Unrotated set did not, were considered cohort views.

After extensive practice, in the *Surprise* phase, subjects were shown all of the objects, both targets and distractors from both the Rotated and Unrotated sets, in a much wider range of probe viewpoints (as well as the viewpoints used in the Practice phase). Otherwise the task remained unchanged, with subjects naming objects or identifying them as distractors.

Subjects were divided into three groups, each of which learned a different set of four target objects – this counterbalancing was done to guard against the idiosyncratic effects of any single object. In each block of trials in the Practice phase the two target objects and the three distractors in the Rotated set appeared at the training viewpoint and at separate rotations of 40°, 70°, and 190° around the *X*, *Y*, and *Z* axes (for a total of ten different viewpoints). The two target objects in the Unrotated set appeared at the training viewpoint only. The two target objects in the Rotated set appeared three times each at the training viewpoint and three times each at the other nine rotations. The two target objects in the Unrotated set appeared 30 times each at the training viewpoint. The three distractors appeared one time each at the same ten viewpoints as the Rotated target objects. A block in the Practice phase consisted of a total of 150 trials, preceded by six preliminary trials.

In each block of trials in the Surprise phase the target objects, whether from the Rotated or Unrotated set, appeared three times each at the training viewpoint and at the 33 viewpoints generated by rotation increments of 30° around the *X*, *Y*, or *Z* axis, while the distractor objects appeared one time each at the same 34 viewpoints. A block in the Surprise phase consisted of a total of 510 trials, preceded by six preliminary trials.

The experiment proceeded for 4 days as follows. On the first day subjects were trained to name the target objects and then ran in two blocks of the Practice phase. On the second and third days subjects ran in four blocks of the Practice phase on each day. On the fourth day, subjects ran in two blocks of the Practice phase and then one block of the Surprise phase. Both the Practice and Surprise phases also shared the following elements: feedback for incorrect responses was provided by a sharp beep; subjects were given a 5 s deadline and failure to respond within this deadline also resulted in a sharp beep; short rests were given to subjects every 50 trials; and, trials within each block were randomly ordered with a different random order for each subject on each block.

2.2. Results and discussion

Mean response times were computed from all correct naming responses collapsed over all stimulus subsets and objects within either the Rotated or Unrotated set; responses for distractors, preliminary trials, and trials where the subject did not respond within a 5 s time limit were discarded.

During the Practice phase naming times for objects in the Rotated set were initially dependent on the distance from the training viewpoint (mean response times were as follows in Block 1, training view: 3101 ms, *X* axis: 3785 ms, *Y* axis: 3428 ms, *Z* axis: 3537 ms; slopes, which measure the putative rate of normalization, were as follows in Block 1, *X* axis: 150°/s, *Y* axis: 343°/s, *Z* axis: 244°/s). With extensive practice, the effect of viewpoint diminished to near equivalent performance at all familiar viewpoints (mean response times in Block 12, training view: 1637 ms, *X* axis: 1714 ms, *Y* axis: 1639 ms, *Z* axis: 1643 ms; slopes in Block 12: *X* axis: 2385°/s, *Y* axis: 707°/s, *Z* axis: 2464°/s). These trends in response times were confirmed by a Block (1 vs. 12) × Viewpoint (40°, 70°, or 190°) ANOVA, where there were reliable main effects of Block, $F(1,11) = 179$, $P < 0.001$, Viewpoint, $F(2,22) = 12.4$, $P < 0.001$, and a reliable interaction, $F(2,22) = 4.31$, $P < 0.05$. Error rates for the Rotated set ranged from 23% in Block 1 to 5% in Block 12. For all blocks of Experiment 1, including Block 13, error rate patterns were consistent with response time patterns. Naming times for objects in the Unrotated set decreased with extensive practice (because these objects appeared at a single viewpoint, the effect of viewpoint could not be assessed during the Practice phase; the mean response times for the single view were 2512 ms in Block 1 and 1411 ms in Block 12). An ANOVA with Block (1 vs. 12) as the only factor revealed that this decrease in response times was reliable, $F(1,11) = 126$, $P < 0.001$. Error rates for the Unrotated set ranged from 4% in Block 1 to 1% in Block 12. Note that the lower error rates obtained for the Unrotated set as compared to the Rotated set are consistent with the fact objects in the Unrotated set appeared only at the canonical training view, while objects in the Rotated set appeared at unfamiliar views where recognition would be expected to be less accurate. Diminished viewpoint dependence with practice replicates Tarr and Pinker (1989) and Tarr (1995).

In the Surprise phase, naming times for the familiar objects in the Rotated set appearing at unfamiliar viewpoints were generally dependent on the distance from the nearest familiar viewpoint (Fig. 3). The two exceptions to this are the familiar viewpoints of 40° and 70° for *X* axis rotations where response times are slower than might be expected for familiar viewpoints; for a similar experiment using the same objects see Tarr (1995). However, naming times for the 40° and 70° *Y* and *Z* axis rotations were just as fast as those for the 10° view. Even given deviations from the predicted pattern, for the objects in the Rotated set regressing response times against distance from the nearest familiar viewpoint resulted in comparable slopes between Block 13 (Fig. 3), where objects were familiar in several views, and Block 1, where objects were familiar in only one view.

Overall, the pattern of viewpoint dependence was systematically related to the distance from the nearest familiar view in both Blocks 1 and 13 and the pattern of response times for the Rotated set again replicates Tarr and Pinker (1989) and Tarr (1995). This result lends credence to the conclusion that extensive practice led subjects to encode viewpoint-specific representations at each practiced viewpoint.

The question is, how did these learned views for the Rotated set influence the recognition of the Unrotated set in these same views? As shown in Fig. 3 both qualitative appearance and quantitative measures indicate that the pattern for the Unrotated set is similar to that obtained for the Rotated set. The similarity between the patterns of response times observed for the two sets may be assessed by correlating response times at each viewpoint for the Rotated set with response times at each viewpoint for the Unrotated set. This analysis revealed reliable correlations for all three axes of rotation: *X* axis: $r(10) = 0.65$, $P < 0.05$, *Y* axis: $r(10) = 0.77$, $P <$

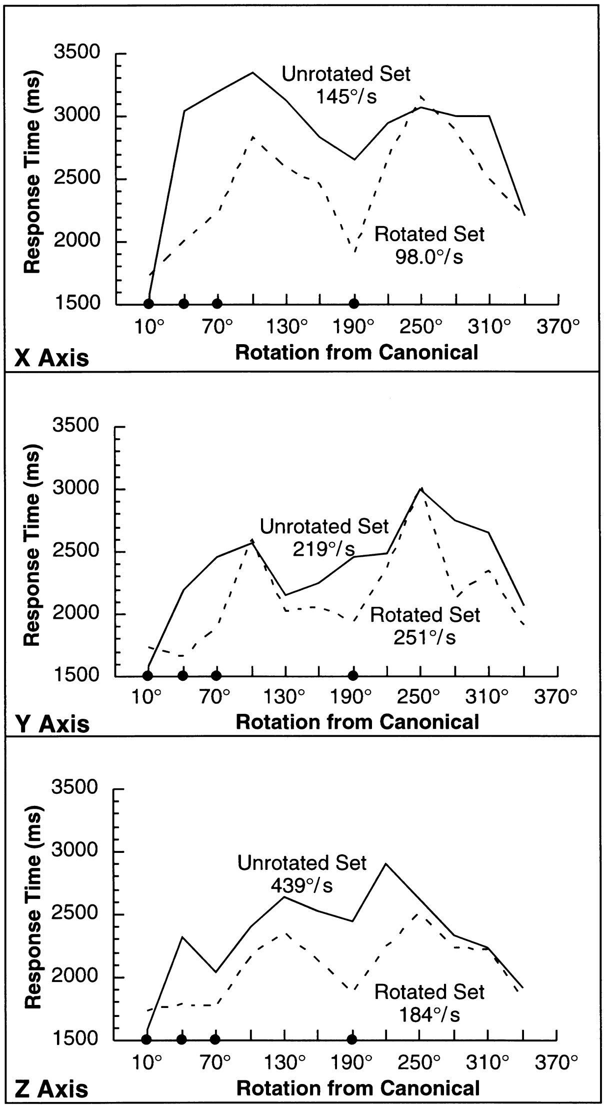

Fig. 3. Experiment 1. Mean response times for the correct identification of named objects for the Rotated and the Unrotated sets as a function of viewpoint and axis of rotation for the Surprise phase. Because rates of rotation vary with axis of rotation (Parsons, 1987) and because subjects appear to rotate through approximately the shortest 3D path (Tarr, 1995), each set of viewpoints around a given axis is displayed separately. Familiar practiced viewpoints (for the Rotated set) are marked with filled circles along the abscissa. The 10° rotation is the canonical training viewpoint and is the same across all three axes. Reported slopes were computed for Block 13 by averaging over all response times for unfamiliar viewpoints equidistant from a familiar viewpoint and then regressing these averaged response times against the distance to the *nearest* familiar viewpoint. For the Rotated sets the familiar viewpoints were defined as those viewpoints actually seen during the Practice phase. For the Unrotated sets the familiar viewpoints were defined as all cohort views.

0.01, *Z* axis: $r(10) = 0.73$, $P < 0.01$. Examining Fig. 3, however, it is also apparent that two patterns emerge depending on which cohort views are considered.

First, for the cohort views of 40° and 70° the predicted transfer from practice with the Rotated set to the Unrotated set does not seem to occur. In other words, response times for the Unrotated set at 40° and 70° are consistent with the pattern expected if only the training view was familiar (e.g. the pattern obtained by Tarr, 1995) – systematically increasing response times with increasing distance from this single familiar view. Why might this be the case? One possibility is that image-based views do not generalize to visually-similar objects. Although this explanation cannot be ruled out, it becomes less plausible when we consider the results for the 190° cohort view and for Experiments 2 and 3. An alternative possibility is that the 10° training view is highly canonical, and as such, subjects never learned (i.e. *X* axis rotations) or transferred (i.e. *Y* and *Z* axis rotations) viewpoint-specific representations for the practiced viewpoints of 40° and 70°. Why might the canonicality of the 10° view produce reduced transfer effects at adjacent viewpoints? It is known that canonical views are often weighted more heavily in determining the most effective match for objects appearing at nearby, albeit familiar, viewpoints (Palmer et al., 1981). There are several reasons for this to be true in Experiment 1. Subjects were taught the names of the target objects by repeatedly seeing only the 10° view. Thus, this view was presented first and more frequently than other views at the beginning of the experiment. Moreover, as illustrated in Fig. 4, this view displays the same surfaces as in the 40° and 70° views. Thus, the additional views provided little new information regarding the appearance of each object and distinct new views would be less likely to be encoded (Tarr, 1995).

Second, for the cohort view of 190° the predicted transfer from practice with the Rotated set to the Unrotated set does occur (Fig. 3). Specifically, when objects in the Unrotated set appeared at or near at the unfamiliar viewpoint of 190°, response times were generally dependent on this viewpoint – one at which objects in the Rotated set were familiar (this is confirmed by the similarity of slopes measured for the Rotated and Unrotated sets). In particular, diminished effects of viewpoint are apparent at 190° for all three axes of rotation – the familiar cohort view furthest from the familiar training viewpoint. If we consider only performance at 190°, the apparent transfer from the Rotated set to the Unrotated set replicates the pattern of viewpoint generalization obtained by Jolicoeur and Milliken (1989) and Murray et al. (1993). As stated earlier, however, these studies failed to probe viewpoints surrounding the cohort views to ascertain whether this transfer was mediated by viewpoint-dependent or viewpoint-invariant object representations (the latter being the usual interpretation). Thus, the pattern of recognition performance at viewpoints where neither objects from the Rotated or the Unrotated sets had previously appeared is particularly diagnostic. As shown in Fig. 3 response times at these viewpoints systematically increased with increasing distance from 190° and the training viewpoints. This increase in response times is not predicted by a viewpoint-invariant account, but is consistent with a viewpoint-dependent image-based account (Bülthoff et al., 1995). The fact that this pattern was obtained for the Unrotated set extends earlier findings of viewpoint dependence, e.g. (Tarr, 1995), indicating that viewpoint-spe-

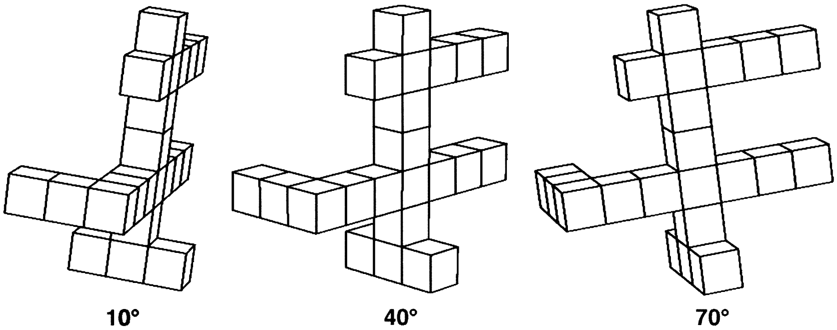

Fig. 4. The 10°, 40° and 70° views for one of the target objects used in Experiment 1. The 10° view was used for teaching subjects the name for each target object and therefore was presented both first and more frequently. Enhancing the canonicality of this view, the 40° and 70° views display the same visible surfaces as the 10° view.

cific representations may generalize to visually-similar objects never actually seen at the familiar views.

One question that arises is why we obtained generalization for the cohort view of 190°, but not for 40° or 70°. We have already discussed reasons why the canonicality of the 10° training view may have ameliorated the influence of adjacent views. Beyond this there may be other factors that facilitated generalization at the 190° view. In particular, geometric and structural similarity between this cohort view and the training view may have allowed better perceptual inferences regarding the appearance of objects in the Unrotated set at this viewpoint. First, consider that the silhouette of the 190° view is almost an exact mirror image of the 10° view (Hayward, 1998). It is well known that observers often show invariance across mirror reflection (Biederman and Cooper, 1991; Cooper et al., 1992). Note that this silhouette similarity did more than simply facilitate recognition of the 190° view – the roughly linear increase in response times with distance from this view indicates that a viewpoint-specific representation was instantiated and served as a target for recognition of adjacent viewpoints. Second, given that some objects actually appeared in the 190° view, subjects had reliable information regarding the orientation of surfaces for objects never seen at that view (since all objects shared the same components). Third, subjects might have been able to use symmetries and other structural regularities within objects to extrapolate or project so-called virtual views at unseen viewpoints. One possible mechanism for the creation of virtual views has been proposed by Poggio and Vetter (Poggio and Vetter, 1992; Vetter et al., 1994). Importantly, all of the objects used in Experiment 1 contained many symmetries

and a highly regular structure. Thus, it may be that image-based generalization between visually-similar objects is only possible or at least more likely when there is supporting geometric and structural information regarding the potential appearance of objects from new viewpoints – a hypothesis we explore in Experiment 2.

Finally, before accepting the proposal the image-based views can generalize between members of a visually-similar class, we should consider an alternative – that subjects learned a common shape-independent viewpoint-specific reference frame for each cohort view. The use of such abstract reference frames would serve equally well for objects in both the Rotated and Unrotated sets and would allow subjects to bypass some of the typically viewpoint-dependent mechanisms required for locating the top, bottom, and major axes of objects prior to recognition (Tarr and Pinker, 1991; Hummel and Biederman, 1992). Indeed, because all of the objects used in Experiment 1 shared a common foot and main axis, we may have made it relatively easy for subjects to learn an abstract reference frame at each practiced viewpoint. Although this explanation cannot be ruled out, there are two points suggesting that subjects did not learn abstract reference frames. First, in Experiment 1 the putative rates of normalization in Block 13 were much slower than those usually associated with top-bottom and axis finding procedures, but were in the range of rates found in other studies where normalization has been the preferred explanation (Parsons, 1987; Tarr and Pinker, 1991; Cohen and Kubovy, 1993; Tarr, 1995). Thus, the recognition of the objects in the Unrotated set is consistent with a dramatic change in performance specifically predicted by normalization to cohort views, rather than the slight dimunition in viewpoint dependency related only to specifically-learned views that might result from bypassing relatively fast axis-finding procedures. Second, Gauthier and Tarr (1997b) found that orientation priming was shape dependent, in that, within a single homogeneous class, orientation priming was larger for those objects that were judged by subjects to be most similar. These results suggest that the representations at cohort views are specific to familiar objects rather than shape-independent reference frames. Finally, Experiments 2 and 3 address this alternative directly, providing evidence for image-based class generalization under conditions where a shape-independent reference frame account cannot hold.

3. Experiment 2

Experiment 1 establishes that viewpoint-specific knowledge about the appearance of objects may transfer to visually-similar cohort objects. This finding provides evidence that frequency of appearance, not just of a given exemplar, but of all members of a perceptually-defined object class, plays an important role in structuring view-based object representations – that is, determining which views are represented in visual memory. In Experiment 1, however, there were frequently-appearing viewpoints where subjects did not appear to learn viewpoint-specific representations – at viewpoints near to the canonical training view (40° and 70°), response times for objects in the Unrotated set increased with distance from the training view. Why might this be the case?

One possibility is that frequency, either of objects or their cohorts, is not the sole factor in determining whether representations are instantiated at particular viewpoints (otherwise diminished effects of viewpoint should occur at all cohort views). In particular, changes in image structure, as determined by the geometry of an object, may account for the presence of a view at 190° and the absence of views at 40° and 70°. Consider that the image structure of objects appearing at these latter two viewpoints will be quite similar to the image structure at the training view (Fig. 4). In contrast, the overall image structure of the same objects appearing at 190° will be quite different as compared to the training view. Thus, views may be stored preferentially in those instances where there are significant qualitative changes in image structure – a hypothesis reminiscent of aspect graph theory (Koenderink, 1987; Effelterre, 1994). Note, however, that the aspect graph approach assumes that views are defined only by qualitative transitions in image structure – here we are suggesting that qualitative variation is simply one factor taken into account in defining views within the representation. Again consider the pattern of results obtained in Experiment 1. The frequent appearance of objects at 40° and 70° was not reason enough to learn representations at these viewpoints given a qualitatively-similar preexisting view at 10° (which was presumably sufficient for recognition of objects appearing at 40° and 70°, albeit only through normalization processes). Consequently cohort views were learned only when the appearance of cohorts coincided with viewpoints that produced drastic changes in image structure, i.e. the 190° view.

Given the above argument, one might expect a difference between depth rotations that produce changes in image structure (rotations around the *X* axis or *Y* axis in Experiment 1) and image-plane rotations that do not produce changes in image structure (rotations around the *Z* axis). Experiment 1, however, does not offer any strong evidence for a difference between these two cases in terms of the particular views that were learned by subjects (nor does Tarr (1995)). In Experiment 2, we wish to consider the possibility that two distinct processes operated in Experiment 1.

First, for any viewpoint of an object, the more often a cohort of visually-similar objects appears at a given viewpoint, the greater the likelihood that this view will be represented in visual memory and will transfer to new exemplars of the class. This is simply an extension of the frequency of appearance principle offered by many researchers (Tarr and Pinker, 1989; Bülthoff and Edelman, 1992; Tarr, 1995) and apparently at work in Experiment 1. Crucially, this type of process is equally applicable to both depth and image-plane rotations. Second, exclusively across rotations in depth, the greater the qualitative dissimilarity between cohort views and actually-seen views, the greater the likelihood that a qualitatively-distinct viewpoint will be represented in visual memory and will transfer to new exemplars of the class. This process is based on the principle that views are organized into subregions (aspects) where the image structure remains relatively stable.

In Experiment 1 it appeared that the first principle was responsible for the transfer observed at the cohort view of 190° and the second principle was responsible for the failure of transfer at other cohort views. This, however, is an admittedly post hoc explanation and for most cases it is difficult to differentiate between these principles in that both may typically facilitate generalization across exemplars. In Experiment 2 we used a design similar to that used in Experiment 1, but we attempted to tease these two factors apart by reducing the effectiveness of the first principle (frequency of appearance) by presenting each exemplar at a different set of viewpoints. This manipulation had the effect of reducing the frequency of any specific viewpoint for the class as a whole. At the same time we hoped to specifically engage the second principle by presenting a large range of viewpoints. In such a case, transfer would be predicted for rotations in depth but not for rotations in the image-plane. Thus, for Experiment 2 we expected that cohort views generated by rotations around either the *X* axis or *Y* axis would transfer to objects in the Unrotated set, but that cohort views generated by rotations around the *Z* axis would not transfer.

3.1. Method

3.1.1. Subjects

Seventeen subjects from the MIT community participated in the experiment for pay. Two subjects were removed from the study because their performance was consistently at chance. All reported normal or corrected to normal vision. None of the subjects who participated in Experiment 2 served as subjects in any other experiment reported in this paper.

3.1.2. Materials

The stimuli used in Experiment 2 were identical to those used in Experiment 1 (see Fig. 1).

3.1.3. Design and procedure

Experiment 2 was quite similar to Experiment 1 with the exception of the specific viewpoints used during the Practice phase and the number of trials in the Surprise phase. The Training phase was identical to that used in Experiment 1. In each block of trials in the Practice phase one target object (A) in the Rotated set appeared at the training viewpoint and at separate rotations of 70° and 190° around the *X*, *Y* and *Z* axes (for a total of seven different viewpoints), while a second target object (B) in the Rotated set appeared at the training viewpoint and at separate rotations of 130° and 250° around the *X*, *Y* and *Z* axes (for a total of seven different viewpoints). The two target objects in the Unrotated set appeared at the training viewpoint only. The two target objects in the Rotated set appeared four times each at the training viewpoint and four times each at the other six rotations, while the two target objects in the Unrotated set appeared 28 times each at the training viewpoint. The two rotated distractors appeared two times at the same seven viewpoints as the rotated targets – one distractor object appearing at each subset of six target viewpoints – while the remaining distractor appeared 14 times at the training viewpoint only. A block in the Practice phase consisted of a total of 154 trials, preceded by six preliminary trials.

In each block of trials in the Surprise phase the target objects, whether from the Rotated or Unrotated set, appeared 12 times each at the training viewpoint and four times each at the 33 viewpoints generated by rotation increments of 30° around the *X*, *Y* or *Z* axis, while the distractor objects appeared six times each at the training viewpoint and two times each at the same 33 viewpoints. A block in the Surprise phase consisted of a total of 792 trials, preceded by six preliminary trials.

3.2. Results and discussion

As in Experiment 1 mean response times were computed from all correct naming responses collapsed over all stimulus subsets and objects within either the Rotated or Unrotated set; responses for distractors, preliminary trials, and trials where the subject did not respond within a 5 s time limit were discarded.

During the Practice phase naming times for objects in the Rotated set were initially dependent on the distance from the training viewpoint (mean response times were as follows in Block 1, for Object A: training view: 2073 ms, *X* axis: 2670 ms, *Y* axis: 2405 ms, *Z* axis: 2474 ms; slopes, which measure the putative rate of normalization, were as follows in Block 1: *X* axis: 683°/s, *Y* axis: 800°/s, *Z* axis: 306°/s; for Object B: training view: 2559 ms, *X* axis: 3125 ms, *Y* axis: 2648 ms, *Z* axis: 2756 ms; slopes: *X* axis: 203°/s, *Y* axis: 1853°/s, *Z* axis: 550°/s. With extensive practice, the effect of viewpoint diminished to near equivalent performance at all familiar viewpoints (mean response times in Block 12, for Object A: training view: 983 ms, *X* axis: 1030 ms, *Y* axis: 1012 ms, *Z* axis: 938 ms; slopes in Block 12: *X* axis: 5043°/s, *Y* axis: 13 569°/s, *Z* axis: −4386°/s; for Object B: training view: 1026 ms, *X* axis: 1048 ms, *Y* axis: 1000 ms, *Z* axis: 1000 ms; slopes: *X* axis: 53 476°/s, *Y* axis: 15 129°/s, *Z* axis: −58 140°/s – such incredibly fast slopes indicate that recognition performance became equivalent at all practiced, familiar views). These trends in response times were confirmed by Block (1 vs. 12) × Viewpoint (Object A: 10°, 70°, or 190°; Object B: 10°, 130°, or 250°) ANOVAs, where for Object A[2] there were reliable main effects of Block, $F(1,13) = 121$, $P < 0.001$, Viewpoint, $F(2,26) = 10.1$, $P < 0.001$, and a reliable interaction, $F(2,26) = 7.52$, $P < 0.005$, and for Object B[3] there was a reliable main effect of Block, $F(1,14) = 289$, $P < 0.001$, a nearly reliable effect of Viewpoint, $F(2,28) = 2.96$, $P = 0.07$, and a marginally reliable interaction, $F(2,28) = 2.36$, $P = 0.11$. Error rates for the Rotated set ranged from 40% for Object A and 43% for Object B in Block 1 to 4.0% for Object A and 5.0% for Object B in Block 12. For all blocks of Experiment 2, including Block 13, error rate patterns were consistent with response time patterns. Naming times for objects in the Unrotated set decreased with extensive practice (because these objects appeared at a single viewpoint, the effect of viewpoint could not be assessed during the Practice phase; the mean response times for the single view were 1880 ms in Block 1 and 855 ms in Block 12). An ANOVA with Block (1 vs. 12) as the only factor revealed that this decrease in response times was reliable, $F(1,14) = 132$, $P < 0.001$. Error rates for the Unrotated set ranged from 9.6% in Block 1 to 2.5% in Block 12. Again, such patterns, particularly diminished viewpoint dependence with practice, replicate the findings of Tarr and Pinker (1989) and Tarr (1995).

In the Surprise phase, naming times for the familiar objects in the Rotated set appearing at unfamiliar viewpoints were generally dependent on the distance from the nearest familiar viewpoint (Fig. 5, left panels). Several results stand out in this

[2] One subject was omitted from the analysis for Object A due to low accuracy rates in Block 1.

[3] Note that for Object B the practice viewpoints were located at 110° (250°) and 130°. As these two viewpoints were almost equidistant from the training viewpoint, practice would not be expected to dramatically change the relationship between response times at these two views.

regard. First, unlike most earlier studies of viewpoint dependence (e.g. Tarr, 1995), different objects were practiced at *different* viewpoints. According to view-based accounts of recognition, this should lead to the instantiation of viewpoint- *and* exemplar-specific representations. Thus, the view-based prediction is that only viewpoints in which a given object was actually seen should show diminished effects of viewpoint and a pattern for nearby viewpoints indicating that the familiar viewpoint was used as a target for normalization. For objects in the Rotated set, this strong prediction holds true. In the Surprise phase, Object A clearly shows systematic viewpoint dependency related to the distance from 10°, 70°, and 190° – the three familiar viewpoints (see Fig. 5). A similar pattern is observed for Object B. There is systematic viewpoint dependency related to the distance from 10°, 130° and 250° – again the three familiar viewpoints. Second, the putative rates of normalization measured by the slopes and the patterns of response times observed in Fig. 5 (left panels) are consistent with results where subjects have learned multiple views of familiar objects. In contrast to Experiment 1, the number of viewpoints per an object and the reasonably wide spacing between practice viewpoints apparently led subjects to instantiate views at every familiar viewpoint. Third, the fact that the slopes for Block 13, where objects were familiar in several views, are comparable to the slopes for Block 1, where objects were familiar in only one view, suggests that similar processes of normalization to a familiar view are operating in both the Practice and the Surprise phases. Overall, the results for the Rotated set again replicate the findings of Tarr and Pinker (1989) and Tarr (1995) regarding the recognition of familiar objects in unfamiliar viewpoints. Thus, we can once more conclude that extensive practice led subjects to encode viewpoint-specific representations at each practiced viewpoint.

As in Experiment 1, the crucial question is how did these learned views for the Rotated set influence the recognition of the Unrotated set in these same views? Inspecting Fig. 5 (right panels) it seems clear that, in contrast to the pattern obtained in Experiment 1, the pattern of response times for the Unrotated set was *not* similar to that obtained for the Rotated set. This was true in terms of the patterns for either of the objects in the Rotated set separately and the pattern expected if all of the cohort views were used as targets for the Unrotated set. In particular, response times for the Unrotated set were not dependent on the distance from the nearest cohort view (when that same view showed evidence of being a learned view for the Rotated set). Notably, the degree to which cohort views had *any* effect on the recognition of the objects in the Unrotated set varied with the axis of rotation. For depth rotations around the *X* axis or *Y* axis, there was little evidence for a specific influence of cohort views, but evidence for a general influence of cohort views. Thus, response times for the Unrotated set for depth rotations followed a pattern suggesting the presence of a single view for which cohort generalization was obtained. In contrast, for image-plane rotations around the *Z* axis, there was *no* evidence for any influence of any cohort views. Thus, response times for the Unrotated set for image-plane rotations followed a pattern suggesting that only the actually-seen canonical training view was used as a target for recognition. These inferences are supported by the correlations between response times at each viewpoint for the Rotated set (both Objects A

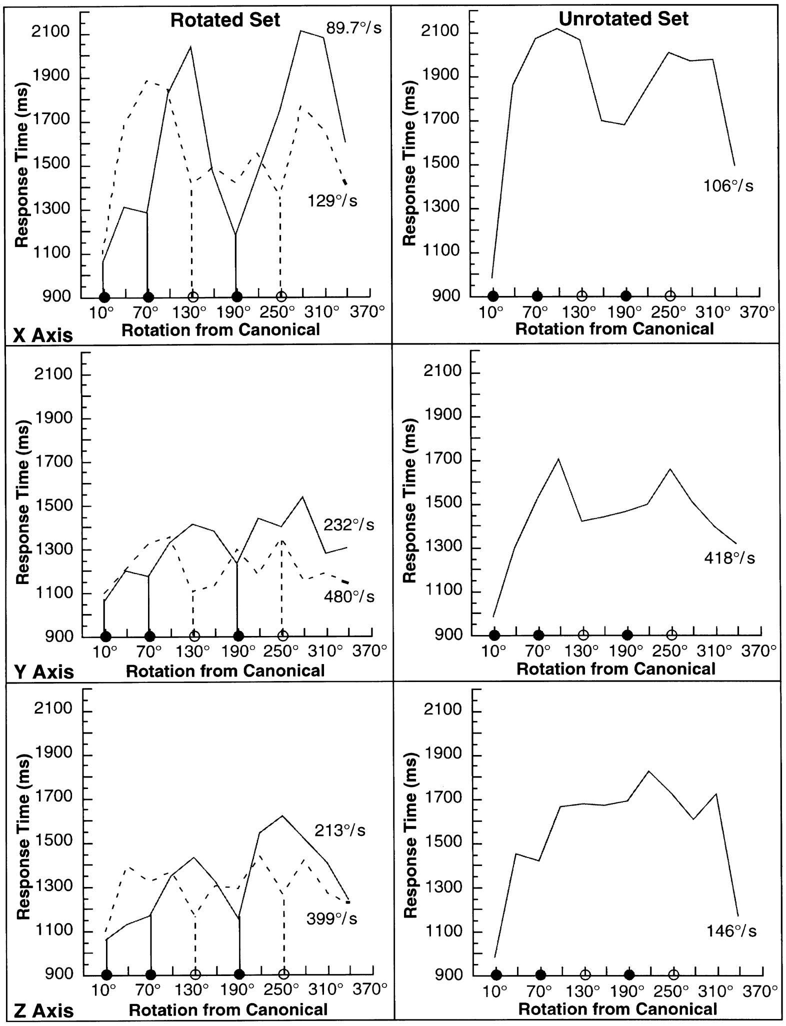

Fig. 5. Experiment 2. Mean response times for the correct identification of named objects for the Rotated and the Unrotated sets as a function of viewpoint and axis of rotation for the Surprise phase. Viewpoints around a given axis were again plotted separately. The left panels show the data for the two named objects in the Rotated set as filled and unfilled circles with the corresponding familiar practice viewpoints marked along the abscissa. The right panels show the data for the named objects in the Unrotated set with the cohort views marked along the abscissa. The 10° rotation is the canonical training viewpoint and is the same across all three axes.

and B) with response times at each viewpoint for the Unrotated set: Object A: X axis: $r(10) = 0.42$, n.s., Y axis: $r(10) = 0.41$, n.s., Z axis: $r(10) = 0.78$, $P < 0.01$; Object B: X axis: $r(10) = 0.65$, $P < 0.05$, Y axis: $r(10) = 0.77$, $P < 0.01$, Z axis: $r(10) = 0.07$, n.s.

We introduced Experiment 2 by suggesting that there are at least two principles that govern when a viewpoint is stored in visual memory. First, frequency of appearance for specific viewpoints has often been shown to prompt the creation of views for the exact exemplars seen in those viewpoints (Bülthoff and Edelman, 1992; Tarr, 1995). Experiment 1 demonstrated that this same principle applies more generally to visually-similar object classes. Second, we hypothesized that differences in image structure that arise with rotations in depth may also prompt the creation of views for both specific exemplars and visually-similar object classes. In Experiment 2 we tested for the existence of this latter principle by making the frequency of appearance principle less potent for the overall object class (by varying the viewpoints in which each exemplar appeared). The results of Experiment 2 suggest that this manipulation was successful. While each exemplar seen at several viewpoints was apparently represented as a set of views located at familiar viewpoints (Fig. 5) left panels, these same cohort views did not transfer to visually-similar objects in either the Rotated or Unrotated sets. This interpretation is supported by the finding that individual objects in the Rotated set showed performance patterns that were viewpoint dependent specifically to the viewpoints in which each object was actually seen – familiar viewpoints for the other member of the Rotated set had little effect. Second, objects in the Unrotated set showed performance patterns that were viewpoint dependent only to the canonical training view or to this view plus a single view approximately 180° from this view – even though, unlike Experiment 1, some of these cohort views were quite far from the training view. Thus, reducing the frequency of any specific viewpoint for the entire class of objects had the effect of disengaging frequency of appearance as a class-general principle for the creation of views in visual memory.

At the same time, this manipulation also had the desired effect of engaging object geometry as the basis for instantiating distinct views in visual memory. This is evidenced by the differences in the patterns obtained for depth rotations and for image-plane rotations for the Unrotated set (Fig. 5) – right panels. Indeed, the strong prediction we made was that familiarity with cohort views would transfer for rotations in depth, but not transfer for rotations in the image plane. Qualitatively, this is the case. For depth rotations there is a clear shift in the pattern of viewpoint dependence relative to the pattern expected if no cohort views were present (that is, performance related only to the distance from the single familiar canonical viewpoint). In contrast, for image-plane rotations there is little shift in the pattern of viewpoint dependence relative to the pattern expected in the absence of cohort views. The most obvious explanation for this set of results is that changes in the image structure influenced the instantiation of views. Specifically, while there were an equal number of cohort views for each axis of rotation, only for depth rotations were the consequent changes in image structure significant enough to warrant the instantiation of new views in visual memory.

Why might the apparent geometric effects for depth rotations manifest themselves only as a single view rather than at each cohort view? Put another way, why did the effects of the cohort views not transfer to each of the specific viewpoints where such a view occurred? As suggested earlier, one possible explanation is that the degree of visual similarity in the image structure at adjacent viewpoints led subjects to ignore some cohort views (Fig. 4). Effectively, the visual system may have computed the qualitative or aspect graph structure for each object or class and then only instantiated views where new aspects were apparent (Koenderink, 1987). This explanation is supported by the observation that some of the cohort views are sufficiently different from the canonical training view to prompt the instantiation of at least one new view. For depth rotations, the presence of the single view where transfer occurred suggests that observers were sensitive only to dramatic changes in qualitative image structure. It should be noted that this result does not support the notion of aspect graphs per se. In particular, current models of how aspect graphs are computed (Freeman and Chakravarty, 1980; Koenderink, 1987) rely on features that are so unstable as to produce intractable numbers of qualitatively-distinct views – many more for the objects used here than is evidenced by the behavioral data.

How then should the specific viewpoint of such a geometrically-defined view be determined? On the one hand, it might be sufficient to store the viewpoint maximally dissimilar from the canonical view; on the other hand, such a view might also be dissimilar from some of the already-seen familiar cohort views. One compromise might be to select a view that is qualitatively distinct from the canonical view, but also easily extrapolated from the canonical views and observed cohort views. Indeed, this latter explanation was suggested in Experiment 1 to account for the finding of cohort generalization for the 190° view, but not for the 40° or 70° views. The same factors that allowed generalization for the 190° case in Experiment 1 are true for the views that show the best transfer in Experiment 2. Adding to these factors, these particular views were intermediate between views at which a subset of the objects were actually seen. Thus, further facilitation may have arisen through view interpolation (Poggio and Edelman, 1990; Bülthoff and Edelman, 1992) or linear combinations of surrounding views (Ullman and Basri, 1991; Ullman, 1998).

In summary, the results of Experiment 2 provide further evidence for viewpoint-specific generalization from some members of a perceptually-defined object class to other members of that same class. Again there is some evidence that familiarity with specific viewpoints transfers to objects never actually seen in those viewpoints. In contrast to Experiment 1, however, generalization in Experiment 2 was also based on how the visible image structure of the objects varied with changes in viewpoint. Thus, there appear to be at least two principles governing how views are created in visual memory: the familiarity of a given viewpoint and the distinctiveness of the image structure for a given viewpoint.

4. Experiment 3

The stimulus set used in Experiments 1 and 2 may be thought of as a homogeneous object class analogous to a set of exemplars drawn from a single basic-level perceptual category (Brown, 1958; Rosch et al., 1976). While Experiment 2 demonstrated that not all familiar viewpoints for such a class generalize to all members of that class, we were also interested in viewpoint-specific generalization in the context of multiple object classes. That is, how would viewpoint-specific familiarity with one exemplar generalize to a second exemplar of the same class under conditions where subjects learned several distinct object categories? The idea that class generalization is mediated by viewpoint-specific image-based views would be supported if familiar views transfer only to objects of the same) perceptual category, thereby indicating that transfer is a function of visual similarity rather than a generic or strategic effect (Gauthier and Tarr, 1997b). Moreover, this transfer effect would be even more striking if it occurred when subjects were discriminating *between* exemplars of a single category. As mentioned earlier, a similar argument has already been made for orientation priming of 2D novel objects (Gauthier and Tarr, 1997b): blocking identification trials by orientation led to significant orientation priming of different objects within the same class but alternating objects between two distinct classes (i.e. qualitatively-different objects) rendered the blocking manipulation ineffective. Note, however, that the orientation priming in Gauthier and Tarr's study was transient in that generalization was tested only over a small number of consecutive trials. Experiment 3 tests whether the same type of class-specific transfer is present for the long-term generalization effects found in Experiments 1 and 2.

A second goal of Experiment 3 was to better measure the variation that is likely to be produced by changes in image structure associated with rotations in depth. As discussed in Experiment 2, the manner in which the visible geometry of an object or class changes is one principle that is likely to govern which views should be represented in visual memory (Freeman and Chakravarty, 1980; Koenderink, 1987; Tarr, 1995). However, to guard against idiosyncratic geometric effects for single objects, Experiments 1 and 2 counterbalanced sets of target objects across subjects. This counterbalancing manipulation meant that results were necessarily averaged over the different sets and, as a consequence, the geometric variation that might arise from specific viewpoints of specific objects would be lost. Thus, an object-specific baseline in the absence of any transfer effects could not be established. Because evidence for transfer between objects is best interpreted in terms of such a baseline, determination of more subtle generalization effects at cohort views was not possible in Experiments 1 and 2. For this reason, Experiment 3 used the same set of objects as targets for all subjects and limited the number of objects in each category to two (one Rotated set object and one Unrotated set object). In this way, the function relating response time to viewpoint in the Transfer condition will reflect more directly the factors of familiarity and object geometry.

We also introduced a second Baseline condition run on a separate group of subjects. In this latter condition, the same objects were trained and practiced only in the canonical training viewpoint and then tested in the same larger set of viewpoints used in the Surprise phase of the Transfer condition. This manipulation afforded us a measure of the variation in response times present when familiar objects were recognized for the first time in unfamiliar viewpoints where there were no familiar or cohort views other than the training view – that is, a condition in which deviations from linearity must be due to object geometry and not familiarity. Note that while this control would be useful in any study that assesses viewpoint dependence across multiple trained viewpoints (e.g. Bülthoff and Edelman, 1992; Humphrey and Khan, 1992), in the present study we omitted it from Experiments 1 and 2 because Tarr (1995) had already collected extensive data on the recognition of the same stimuli across viewpoints under various training conditions. In contrast, the stimuli used in Experiment 3 had not previously been used in any such experiment.

4.1. Method

4.1.1. Subjects

Thirty-four undergraduate students enrolled in Introduction to Psychology at Yale University were given course credit in return for their participation – because of extremely high error rates and excessive variation across different viewpoints in the final test phase of each condition, six subjects were excluded from excluded from the study, leaving twenty-eight subjects. All reported normal or corrected to normal vision. Fourteen of the subjects were run in the Transfer condition and 14 of the subjects were run in the Baseline condition. None of the subjects who participated in Experiment 3 served as subjects in any other experiment reported in this paper.

4.1.2. Materials

Twelve computer-generated 3D objects were created on a Macintosh computer using 3D modeling software (Alias Research, Toronto, Canada). The objects are shown in Fig. 6 in their arbitrarily designated canonical view (0° – leftmost column). There were six targets and six distractors, each group including three pairs of objects sharing the same central part, but with a slightly different arrangement of the smaller attached parts – it was assumed that subjects would treat objects sharing similarly-shaped central parts as members of the same distinct perceptual category (Tversky and Hemenway, 1984; Gauthier and Tarr, 1997a). The targets were given arbitrary names such as ‘Kip’, ‘Kal’, or ‘Mar’. Images of targets and distractors were generated from 12 viewpoints (every 30° around the vertical/*Y* axis). Photo-realistic rendering of these images was done with 24-bit color and then each image was reduced to a common 8-bit palette using Debabilizer (Equilibrium, CA, USA). All objects were colored the same orange-ocher hue. They were presented centered on the screen against a white background. Stimuli were approximately 6.5 cm × 6.5 cm and subjects sat about 60 cm from the screen, yielding a display area subtending approximately 6.2° × 6.2° of visual angle. The experiment was run on an Apple Macintosh LC 475 equipped with a Trinitron 13 inch color monitor with a resolution of 640 × 480 pixels (72 dpi).

4.1.3. Design and procedure

Experiment 3 used the same Training-Practice-Surprise phase sequence used in Experiments 1 and 2. As before, the experiment was spread over 4 days, with the Training phase and two blocks of the Practice phase on day 1, four blocks of the Practice phase on days 2 and 3, and two blocks of the Practice phase and one block of the Surprise phase on day 4. In terms of presentation times, response deadlines, feedback and randomization of trials, Experiment 3 used the same procedures as used in Experiments 1 and 2.

The Training phase consisted of three parts and was identical for the Transfer and Baseline conditions. First, each target object was presented on the screen in its canonical training viewpoint (0°, Fig. 6) for 5 s with its associated name. Subjects were instructed to simply study each object and learn its name. Second, the six target objects were shown four times each with their names, for 5 s each, and subjects were required to press the key labeled with the appropriate name for each object. An incorrect key press resulted in a beep. Third, subjects ran in 36 randomly ordered naming trials with each target being shown a total of six times without its name. Again subjects were required to press the correct key within 5 s or a beep would result.

In the Practice phase subjects practiced recognizing the objects from a small number of viewpoints generated by rotations in depth around the vertical axis (Fig. 6). In the Transfer condition the Rotated set named target objects (one of each pair – $A_2 \rightarrow C_2$) were presented at 10° and at one other viewpoint, either 60°, 150° or 240°. The Unrotated set named target objects ($A_1 \rightarrow C_1$) appeared as

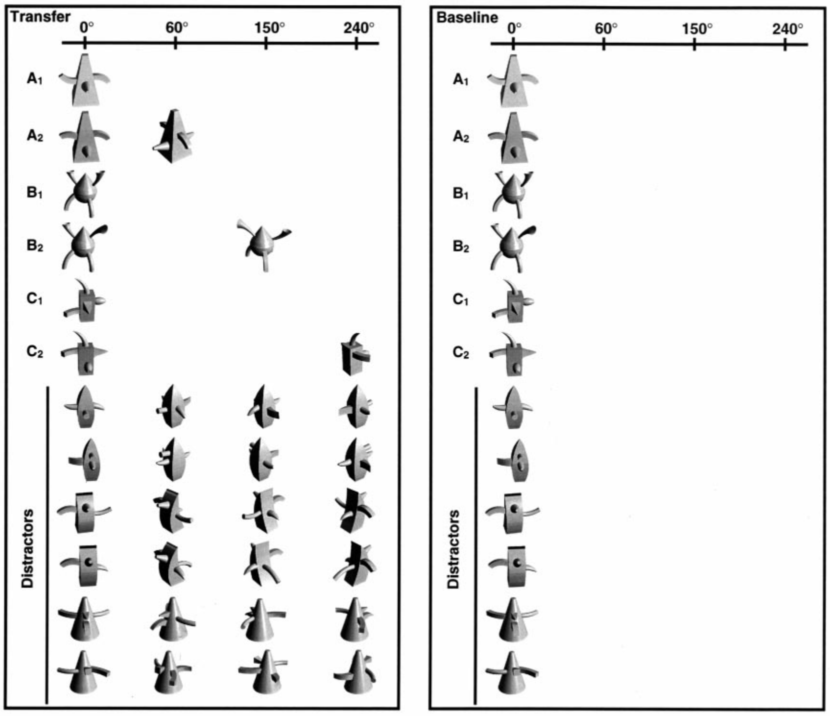

Fig. 6. The set of novel 3D objects used as stimuli in Experiment 3. Both named target objects ($A_1 \rightarrow C_1$, $A_2 \rightarrow C_2$) and unnamed distractor objects are shown in their arbitrarily defined canonical viewpoint (0°). Objects are also shown in the other viewpoints used during the practice phase. The left panel shows the viewpoints used for the Transfer condition and the right panel shows the viewpoints used for the Baseline condition.

frequently as the Rotated set, but only at the canonical training viewpoint of 10°. All six of the distractor objects appeared in all practiced viewpoints. The three target objects in the Rotated set appeared six times each at the training viewpoint and six times each at the other viewpoint. The three target objects in the Unrotated set appeared 12 times each at the training viewpoint. The six distractors appeared once each at the four viewpoints used individually for the objects in the Rotated set. A block in the Practice phase consisted of a total of 96 trials (75% targets/25% distractors). In the Baseline condition all of the named target objects and the distractors appeared only at the canonical training viewpoint (0°). Target objects were shown 12 times each and distractors were shown four times each. Note that the objects in the Unrotated set of the Transfer condition ($A_1 \rightarrow C_1$) appeared in the identical) set of viewpoints repeated the same number of times in the Baseline condition. Thus, the only difference for these objects (and between the Transfer and Baseline conditions) was whether or not their cohorts ($A_2 \rightarrow C_2$) appeared at other viewpoints.

In each block of trials in the Surprise phase for both the Transfer and Baseline conditions the target objects, regardless of the viewpoints shown during the Practice phase, appeared six times each at the training viewpoint and the 11 viewpoints generated by rotation increments of 30° around the *Y* axis. The distractor objects appeared two times each at the same 12 viewpoints. A block in the Surprise phase consisted of a total of 576 trials (75% targets/25% distractors). Note that the Surprise phase for the Transfer and Baseline conditions were identical). Thus, any differences in the pattern observed for target objects must be attributed to either direct or cohort familiarity with viewpoints seen during the Practice phase.

4.2. Results and discussion

Mean response times were computed from all correct naming responses collapsed over subjects; responses for distractors and trials where the subject did not respond within a 5 s time limit were discarded.

During the Practice phase in the Transfer condition naming times for objects in the Rotated set were initially dependent on the distance from the training viewpoint for two of the three objects (mean response times were as follows in Block 1, Object A_2: 1866 ms, Object B_2: 1691 ms, Object C_2: 2291 ms; slopes were as follows in Block 1, Object A_2: 71.6°/s, Object B_2: 641°/s, Object C_2: 157°/s). Note that an exception to the typical pattern of viewpoint dependence was obtained for Object B_2 – response times reflected much greater viewpoint invariance. Object B_2 was highly dissimilar from the other objects and showed the same parts attached to its top regardless of any rotation around the vertical axis – apparently this allowed subjects to use distinctive features to recognize the object (Tarr and Bülthoff, 1995). Given this immediate invariance it was not possible to obtain any evidence for class generalization for Object B_2 (in that no effects of viewpoint were expected for the Surprise phase).

With extensive practice, the effect of viewpoint diminished to near equivalent performance at both familiar viewpoints for Objects A_2 and C_2 (mean response times in Block 12, Object A_2: 750 ms, Object B_2: 743 ms, Object C_2: 774 ms; slopes in Block 12, Object A_2: –1795°/s, Object B_2: –15 267°/s, Object C_2: –11 779°/s). These trends in response times were confirmed by a Block (1 vs. 12) × Viewpoint (60°, 150° or 240°, depending on the object) ANOVAs. For Object A_2 there was a reliable main effect of Block, $F(1,11) = 39.2$, $P < 0.001$, Viewpoint, $F(1,11) = 9.72$, $P < 0.01$, and a reliable interaction, $F(1,11) = 14.6$, $P < 0.005$ (two subjects were excluded from this analysis because they had no correct responses in the Block 1 × 60° cell). For Object B_2 there was a reliable main effect of Block, $F(1,13) = 47.2$, $P < 0.001$, but no reliable main effect of Viewpoint, $F(1,13) = 2.63$, n.s., or interaction, $F(1,13) = 1.70$, n.s. For Object C_2 there was a reliable main effect of Block, $F(1,12) = 52.0$, $P < 0.001$, Viewpoint, $F(1,12) = 4.62$, $P < 0.05$, and a near-reliable interaction, $F(1,12) = 4.07$, $P = 0.07$ (one subject was excluded from this analysis because they had no correct responses in the Block 1 × 240° cell). Error rates for the Rotated set ranged from 24.4% for Object A_2, 17.3% for Object B_2, and 25.6% for Object C_2 in Block 1 to 0.60% for Object A_2, 1.79% for Object B_2, and 2.38% for Object C_2 in Block 12. For all blocks of Experiment 3, including Block 13, error rate patterns were consistent with response time patterns. Naming times for objects in the Unrotated set decreased with extensive practice (because these objects appeared at a single viewpoint, the effect of viewpoint could not be assessed during the Practice phase; the mean response times for the single view were 1300 ms for Object A_1, 1, 448 ms for Object B_1, and 1548 ms for Object C_1 in Block 1 and 748 ms for Object A_1, 741 ms for Object B_1, and 758 ms for Object C_1 in Block 12). ANOVAs with Block (1 vs. 12) as the only factor revealed that this decrease in response times was reliable for all three objects, $F(1,13) = 11.5$, $P < 0.005$, for Object A_1, $F(1,13) = 49.7$, $P < 0.001$, for Object B_1, and $F(1,13) = 52.2$, $P < 0.001$, for Object C_1. Error rates for the Unrotated set ranged from 11.9% for Object A_1, 14.9% for Object B_1, and 7.74% for Object C_1 in Block 1 to 0.60% for Object A_1, 1.19% for Object B_1, and 3.57% for Object C_1 in Block 12. Error rate patterns were always consistent with response time patterns.

During the Practice phase in the Baseline condition naming times for objects decreased with extensive practice. For individual objects response times and errors were quite similar to those obtained in the Transfer condition. Overall, for objects that were members of the Rotated set in the Transfer condition, in the Baseline condition[4] mean response times were 1758 ms for Block 1 and 1098 ms for Block 12 and error rates were 24.6% for Block 1 and 1.98% for Block 12. For objects that were members of the Unrotated set in the Transfer condition, in the Baseline condition mean response times were 1528 ms for Block 1 and 1198 ms for Block 12 and error rates were 17.1% for Block 1 and 2.98% for Block 12.

In the Transfer condition of the Surprise phase, naming times for the familiar objects in the Rotated set appearing at unfamiliar viewpoints were generally dependent on the distance from the nearest familiar viewpoint (Fig. 7). However, an inspection of the response times graphs makes it clear that there were also instances where viewpoint dependency occurred independently of viewpoint familiarity. This

[4] Due to a computer error one subject was not included in the analysis for the Practice phase of the Baseline condition.

can be seen for Object A_2 – there is a unexpected dimunition in response times around the 180° viewpoint. The possible reasons for this pattern may be assessed by comparing the results of the Transfer condition to those of the Baseline condition. For the objects in the Rotated set, the difference between these conditions is that the objects were actually practiced in the cohort view in the Transfer condition, therefore subjects had direct familiarity with these viewpoints. Thus, any apparent viewpoint dependency related to familiarity would be expected to be absent for the Baseline condition.

The fact that familiarity did alter performance at familiar viewpoints may be verified statistically by comparing the response time for a practiced familiar viewpoint for a given object to the response time for that same viewpoint when it is unfamiliar for the same object. This amounts to a *t*-test between the Transfer and Baseline conditions for each object in the Rotated set for its unique familiar viewpoint. These analyses revealed reliable differences for Object A_2, $t(26) = 3.51$, $P < 0.005$, and Object C_2, $t(26) = 2.62$, $P < 0.01$, as well as a near-reliable difference for Object B_2, $t(26) = 1.82$, $P = 0.08$ (suggesting that some viewpoint-specific learning was occurring despite the relative lack of viewpoint dependence during the Practice phase). Alternatively, viewpoint dependency related to properties intrinsic to the stimuli, e.g. object geometry, would be expected to be present for the Baseline condition. Such is the case for the patterns obtained for Object A_2 (and to a lesser extent, Object C_2) – there is a clear parallel between the response times for the Transfer and Baseline conditions for this object. Notably, this unexpected viewpoint dependency was centered around 180° – the most obvious virtual view in terms of both a simple mirror-reflection in the silhouette (Hayward, 1998) and the symmetries present in each object. As discussed earlier, Poggio and Vetter (1992) have proposed a mechanism by which such views may be inferred for symmetrical objects (for related work see Vetter et al., 1994; Logothetis and Pauls, 1995). Our results seem to indicate that this effect is not limited to perfectly symmetrical objects. These observations are reflected by the fact that the slopes measured for Block 13 do not capture the pattern of viewpoint dependence seen in the response time graphs shown in Fig. 7.

As in the previous experiments, the crucial question is how did the learned views for Objects A_2 and C_2 of the Rotated set influence the recognition of Objects A_1 and C_1 of the Unrotated set in these same views? Inspecting Fig. 8 it appears that the pattern of response times for the Unrotated set was sometimes similar to that obtained for the Rotated set. As stated earlier, the exception is Object B_1 – the post hoc explanation for this being the near immediate viewpoint invariance observed in Block 1 for its cohort, Object B_2, and now observed for Object B_1. Given the absence of any significant viewpoint dependency, it is impossible to assess whether transfer occurred between the Rotated and Unrotated sets. Indeed, there was no reliable difference between the Transfer and Baseline conditions for Object B_1 at the cohort view, $t(26) < 1$. Given this result, Objects B_1 and B_2 will not be included in any further analyses or discussion. For the remaining two objects in the Unrotated set, there does appear to be a shift in the pattern of response times at the cohort view and viewpoints nearby this view. This can be clearly seen in Fig. 8

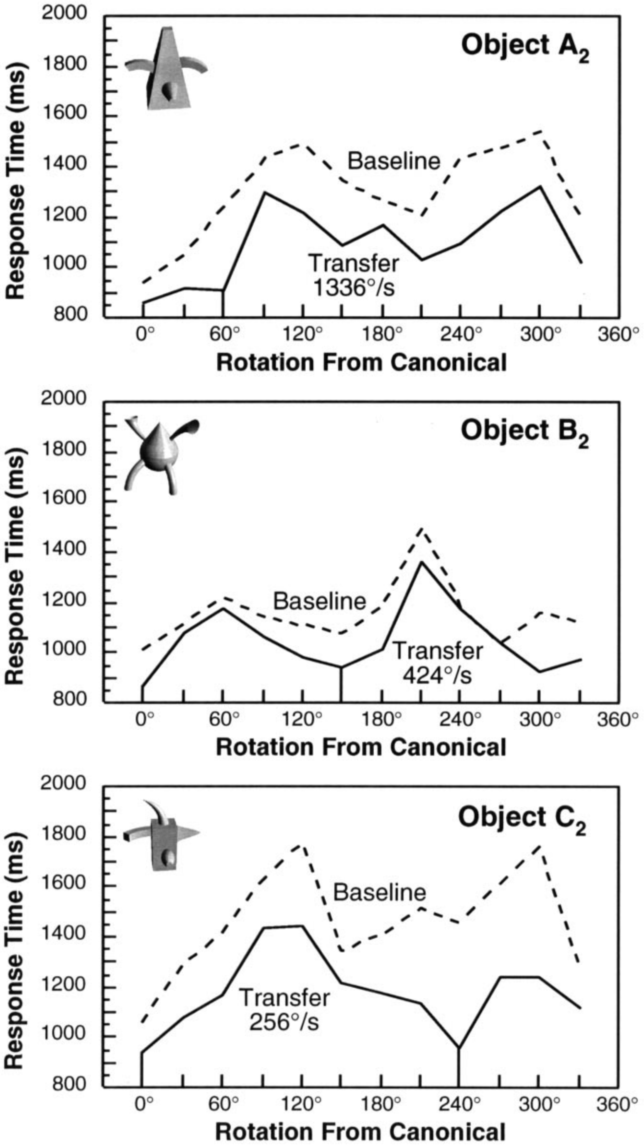

Fig. 7. Experiment 3, Rotated set. Mean response times for correct identification of target objects that were familiar at two viewpoints ($A_2 \rightarrow C_2$) in the Transfer condition as a function of viewpoint for the Surprise phase. The solid line plots the results from the Transfer condition and the dashed line plots the results from the Baseline condition for the same objects. Note that vertical lines mark the viewpoints at which the objects in the Rotated set were *actually* studied during the Practice phase of the Transfer condition; only the 0° viewpoint was used in the Baseline condition. Slopes were not computed for the Baseline condition because of the clear influence of virtual views – an effect not predicted prior to the experiment. Thus, any computation of the putative rates of normalization would be based on post hoc assumptions about the targets of normalization.

by comparing the functions for the Unrotated set objects in the Transfer and Baseline conditions.

It should be emphasized that there is no difference in the *direct* experience subjects received with the objects in the Unrotated set between these two conditions. Therefore, any difference in the response time patterns must be attributable to contextual differences – that is, the cohort views at which objects in the Rotated set appeared. For Object A_1 this shift can be observed in three ways: the lower response time at the cohort view between the Transfer and Baseline conditions, $t(26) = 2.79$, $P < 0.01$; the lower response times at viewpoints adjacent to the cohort view; and, crucially, the shift in the location of the peak denoting the midpoint between views used as targets of normalization. For Object A_1 the peak moves to between 60° and 180° for the Transfer condition. For Object C_1 this shift can be similarly observed: there is a lower response time at the cohort view between the Transfer and Baseline conditions, $t(26) = 1.88$, $P = 0.07$; the dramatically lower response times at viewpoints adjacent to the cohort view; and, crucially, the absence of a peak denoting the midpoint between views used as targets of normalization. For Object C_1 the peak shifts from being between 180° (the mirror-image silhouette and the geometrically-defined virtual view) and 360° for the Baseline condition to being entirely absent for the Transfer condition.

To summarize, Experiment 3 had two goals. First, we were interested in viewpoint-specific generalization in the context of multiple object classes. Here we found that viewpoint-specific familiarity with one exemplar of a perceptually-defined class generalized *only* to other exemplars of that same class and not to exemplars of other classes. Indeed, this class generalization occurred despite the fact that subjects were discriminating between members of a class. This is evidenced by the patterns of performance obtained for Objects A_1 and C_1 in the Transfer condition relative to the patterns obtained for the same objects in the Baseline condition (the only difference between conditions being the viewpoints in which *other* members of each class appeared). Second, we were interested in comparing object- and class-specific familiarity effects to class-general geometric effects as caused by variations in image structure. We found that there were class-general viewpoint dependencies – in particular, for the silhouette mirror-image and at virtual views where symmetry relationships in the bounding contours provided information about the appearance of objects not actually seen at those viewpoints (Hayward, 1998). This is evidenced by the similarity in the patterns of performance between the same objects in the Transfer and Baseline conditions at non-familiar/cohort views (specifically at viewpoints where previous studies suggest that there should be much larger costs for recognition, (Tarr, 1995). Overall, such results lend further support to the hypothesis that image-based representations may support generalization across members of a class. These results, however, also indicate that transfer effects may be relatively subtle and expressed differently depending on the geometry of an object as well as the specific image structure at familiar views. Thus, some of the evidence garnered in Experiment 3 would most probably have been undetectable were results averaged over several different target objects as in Experiments 1 and 2.

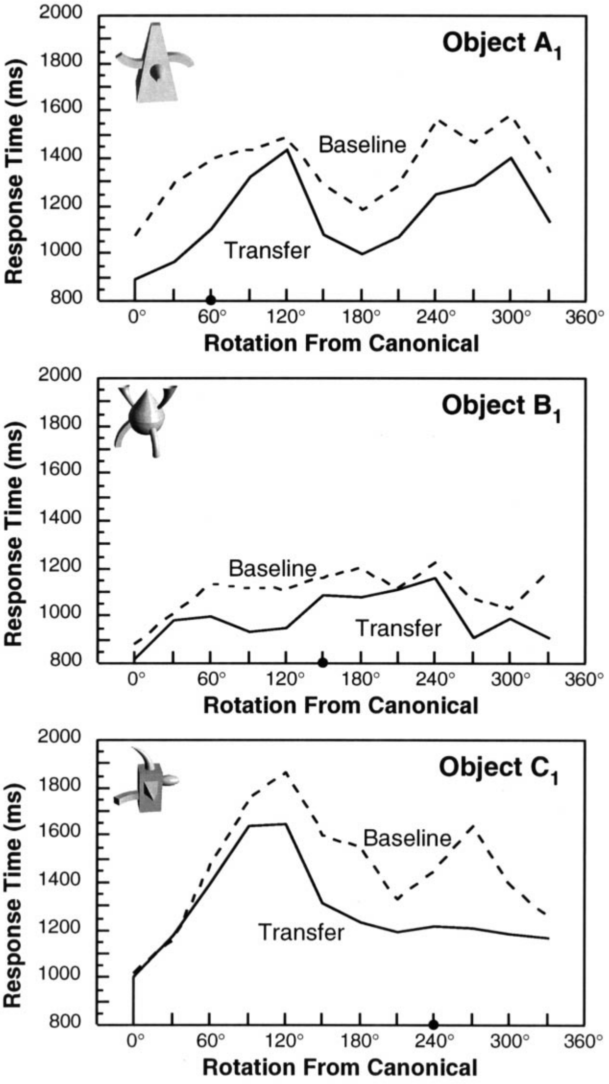

Fig. 8. Experiment 3, Unrotated set. Mean response times for correct identification of target objects that were familiar at only one viewpoint ($A_1 \rightarrow C_1$) in the Transfer and the Baseline conditions as a function of viewpoint for the Surprise phase. The solid line plots the results from the Transfer condition and the dashed line plots the results from the Baseline condition for the same objects. Note that vertical lines mark the viewpoint at which the objects were *actually* studied during the Practice phase of both conditions; the black dots mark the viewpoints at which objects in the Rotated set were studied during the Transfer condition. Note that there was no difference between the Transfer and Baseline conditions for the objects in the Unrotated set – any differences must be attributed to differences in familiarity for views of the Rotated set. Slopes were not computed for the Unrotated set and the Baseline condition because of the clear influence of virtual views – an effect not predicted prior to the experiment. Thus, any computation of the putative rates of normalization would be based on post hoc assumptions about the targets of normalization.

5. General discussion

We began this paper by asking whether image-based recognition mechanisms are capable of generalizing from known instances of a class to unknown instances of that same class. Because earlier image-based mechanisms such as those proposed by Poggio and Edelman (1990) have been associated with rigid templates (Biederman and Gerhardstein, 1995) it has often been assumed that they are incapable of such class generalization. Given that basic-level or categorical recognition is an important element of everyday object recognition, it is critical that any theory of visual recognition exhibit some stability across object classes (Marr and Nishihara, 1978). Indeed, several computational models of image-based recognition tacitly acknowledge this fact in their attempts to develop object representations that generalize across members of a given class. For example, the model proposed by Edelman and Weinshall (1991) used blurred template matching – a process that presumably would allow for greater generalization across members of a class (although with a some concomitant loss of sensitivity). More recently, Edelman (1995) (see also Edelman et al., 1996) has proposed an approach to visual representation in which objects are stored in terms of their relative similarity in a high-dimensional space. Importantly, this representation allows for dynamic access to task-relevant features, thereby supporting both categorical and exemplar-specific recognition within a single system. Some experimental implications of this approach for class-based generalization in face recognition have been explored by O'Toole et al. (1998). Finally, also in the domain of face recognition, Beymer and Poggio (1996) describe a model that uses a vector representation of images and computes a dense correspondence, e.g. visual similarity, between each image in the database. They suggest that this process allows the learning of a *flexible* exemplar-based model for a class of objects. Moreover, this flexible model specifically allows the generation of new virtual views for an object from a single example view, represented as a 2D shape vector, if appropriate prototypical views of other objects in the same class are available (Beymer and Poggio, 1996). Here we investigated whether a similar class-generalization mechanism is used in human object recognition. The results of three experiments indicate that viewpoint-specific representations are *sensitive* enough to support discrimination between visually-similar exemplars, yet *stable* enough to generalize to visually-similar exemplars (transfer views). More specifically, we found that:

- Experience with an object at a given viewpoint transfers to visually-similar objects, albeit in a viewpoint-dependent manner. Thus, the representations or virtual views instantiated for new objects appear to share a status similar to that of viewpoint-specific image-based representations arising from extensive practice (Experiment 1).
- Viewpoint-specific image-based representations are encoded according to at least two principles: the frequency with which objects of a given class are seen at specific viewpoints and the distinctiveness of these views in terms of their geometrical image structure (Experiment 2).

- Transfer between image-based views is class specific (as defined by the visual similarity for a set of objects) in the sense that generalization across viewpoint-specific representations occurs only between objects of a visually homogeneous class (Experiment 3).

Overall, these results are consistent with the hypothesis that human perceivers learn viewpoint-specific representations based on: the frequency with which a specific object appears at different viewpoints; the frequency with which visually-similar objects appear at different viewpoints; and, the distinctiveness of the image structure of different viewpoints relative to other known viewpoints. In total, these findings represent a significant extension to the image-based approach to object recognition.

Although both viewpoint-dependent performance (Jolicoeur, 1985) and diminished viewpoint dependency with increasing familiarity (Tarr and Pinker, 1989; Bülthoff and Edelman, 1992; Tarr, 1995) have been reported previously, the present experiments provide a demonstration that experience with an object’s visually-similar cohorts can facilitate its recognition at novel viewpoints – presumably through the instantiation of image-based views. Moreover, by probing recognition performance at unfamiliar viewpoints adjacent to such views, our results demonstrate that this class generalization is mediated by viewpoint-specific representations.

Earlier evidence for class transfer between objects seen at a given view and objects not seen at that view (Jolicoeur and Milliken, 1989; Murray et al., 1993) can be reinterpreted based on our present results. Specifically, such findings were typically taken as support for the existence of viewpoint-invariant object representations. Our results, however, suggest that class transfer can occur when subjects learn *viewpoint-specific* representations that have some visual similarity or visual feature overlap with new objects subsequently presented at the familiar viewpoints. One caveat worth mentioning regarding our interpretation of these earlier results is that both studies found transfer between familiar objects that, for the most part, were members of different basic-level classes. Thus, the visual similarity between objects actually seen at a given orientation and the transfer objects was presumably somewhat less than that between the novel objects used in the present experiments. On the other hand, several factors may have contributed to obtaining viewpoint-specific transfer even with low object similarity. First, all of the familiar objects used in these studies had a canonical upright orientation relative to gravity. Second, a constant test orientation was used for all of the objects. Third, rotations were always in the picture plane. As a consequence of these factors, the intrinsic axes of the objects were similar across classes, the relative change in the tops and bottoms of the objects was consistent across the experiment, and the image structure of objects from familiar to unfamiliar orientations remained unchanged. Taken together, these conditions may have resulted in transfer orientations in which the appearance of unfamiliar objects was highly predictable. Unfortunately, in these earlier studies (Jolicoeur and Milliken, 1989; Murray et al., 1993) the possibility that transfer was mediated by viewpoint-specific representations was never considered and, as a consequence, the view specificity of the facilitation obtained for objects seen for the first time at new viewpoints was not tested. Obviously, further investigation is necessary to establish whether our account of image-based transfer holds for the experimental conditions used by other researchers.

Interestingly, our results also allow us to reinterpret a classic finding of Bartram (1974). Bartram tested naming performance across blocks of trials in which pictures of objects could be the same, pictures of objects could vary in viewpoint, or pictures of objects could be different objects with the same names as previously named objects. He found strong practice effects in all three conditions, including instances where subjects used the same names for new pictures. Bartram also investigated what happened when subjects were switched from one condition to another. Here he found that there was good transfer from named objects in one view to the same objects in new views, but little transfer from named objects to new objects with the same names. His interpretation of these results was that memory for pictures includes both visual 2D (stimulus) and visual 3D (object) codes. Interestingly, Bartram observed that his results were consistent with an exclusively 2D code if transfer was a function of the extent to which physical features present in one picture overlap with features in other views of the same object. Moreover, he presented some data supporting this hypothesis, pointing out that naming latencies for new viewpoints were more variable than for familiar viewpoints and that this difference may have been related to the fact that the degree of overlap between different spatial viewpoints was varied from almost complete (45° rotation) to minimal (180° rotation). Thus, Bartram's results are consistent with modern theories of view normalization (Tarr and Pinker, 1989; Poggio and Edelman, 1990; Bülthoff and Edelman, 1992).

Regarding same-name/different-picture manipulations, Bartram concludes that the fact that subjects showed practice effects for continuously naming new objects with the same names as previously named objects, but did not show transfer from naming the same objects several times to new objects with the same names is evidence for the presence of a semantic code. He argues that any practice advantage obtained for *repeatedly* applying the same name to different objects must be semantically mediated because the visual codes that would produce such practice effects would, according to his reasoning, also result in transfer when the same object is named for only one or two *presentations* and a new object with the same name is then shown. Based, however, on the recent results of Gauthier and Tarr (1997b), class transfer between different exemplars of an object class may require many exposures (in order to build up sufficient activation in the recognition network – see the discussion below) and may be highly viewpoint specific (Bartram is not clear about whether different same-name objects are shown at the same viewpoints used for earlier exemplars of the class). Thus, Bartram may have obtained practice effects when subjects continuously named new exemplars of a class because class-general activation accumulated over many trials, but failed to obtain transfer effects when subjects named new exemplars after only a few trials because of insufficient activation and because the viewpoints for different objects were not held constant. Indeed, the class transfer reported in this paper is the end-result of many repetitions of the same objects in the same viewpoints. By this account, Bartram's experimental findings are consistent with a model of object recognition in which viewpoint-specific representations mediate both view and class generalization based on visual similarity.

5.1. An image-based network for class recognition

In Section 1 we proposed that a network of linked image-based viewpoint-specific representations could support both subordinate-level discriminations *and* basic-level generalizations. Fig. 9 illustrates a simple conceptualization of such a network that is composed of units (possibly ensembles of neurons) that represent particular exemplars of objects at familiar viewpoints. The key idea of this model is that averaging across different subpools of these units could yield descriptions that would be well suited for basic-level or for subordinate-level recognition tasks. Upon presentation of a stimulus, viewpoint-specific units coding for all visually-similar objects would be activated in proportion to the degree of image-based similarity or feature overlap (for specific models of how such high-dimensional feature spaces might be created see Edelman, 1995; Edelman et al., 1996). For example, when shown a front view of a Victorian chair, the unit coding for the front view of that chair *and* the units coding for other views of the same chair would be activated. Activation related to the degree of view similarity is similar to the neural model of viewpoint-dependent recognition proposed by Perrett et al. (1998). Additionally, because objects of the same class share a configuration of features, units coding for the front view of *other exemplars* of the class of chairs would also be activated. With a large number of exemplars of the same class stored in such a network, clusters of visual similarity would result in a coarse description of the object class, in other words, a visual representation of the basic level (Edelman (1995) proposes a similar population response for representing the basic level). Note that this model may also help to account for the fact that visually atypical members of a class are often named with greater specificity or at what is sometimes referred to as the entry level (Jolicoeur et al., 1984). For example, penguins are typically first identified as penguins rather than birds (the putative basic-level category). Due to the relatively low visual similarity between most birds and penguins, exemplars of penguins will not be included within the cluster of units that defines the category 'birds'. Moreover, when an image of a penguin is encountered it will almost exclusively activate units coding for views of penguins – thus, access to the category penguin (rather than bird) is immediate. In contrast, when subordinate-level or exemplar-specific discrimination is required, a cluster of visually-similar views could eventually arrive at a state in which the most appropriate view reaches a threshold and 'wins' over the other exemplar-specific representations. Thus, a subordinate-level task would set a relatively high threshold of pooled activation as compared to a basic-level task.

5.2. Neural correlates

In this special issue, Perrett et al. (1998) present an elegant model of how cumulative evidence from cells tuned to image-based features can provide the information necessary for recognition *and* account for behavioral effects such as viewpoint dependency. They propose an approach in which the neural response to the presentation of a complete object can lead to faster accumulation of evidence as compared to the accumulation of evidence for individual parts of the same object. This behavior holds even if there are more cells tuned to individual parts than the whole: presentation of the complete object leads to the activation of more cells that contribute to the recognition of the object. Importantly, the model presented by Perrett et al. makes use of the broad tuning across viewpoints, typical of neurons responsive to body parts or other objects (Logothetis et al., 1995), to explain how a population of neurons can respond, albeit with a cost in time, to novel views of familiar objects. There is evidence that visual neurons in the temporal cortex are also broadly tuned to overall object similarity, that is, cells tend to prefer visually-similar pictures (Miyashita et al., 1993). Therefore, Perrett et al.'s model might be extended to account for transfer across different exemplars of a homogeneous class. What is required is that neurons coding for an object's visually-similar cohorts also contribute to the recognition of that object (as suggested in Fig. 9). Consistent with Perrett et al.'s account, we propose that the greater the number of visually-similar exemplars encoded at a given viewpoint the greater the class transfer expected for that particular view (although here we were able to obtain this transfer with only two known exemplars per class – perhaps because of the clearly restricted nature of the class in the context of the experiment). We also suggest that increased variation in the views experienced for a particular class will facilitate transfer by geometrical interpolation (Bülthoff and Edelman, 1992; Librande, 1992; Poggio and Vetter, 1992) by virtue of a large diversity of neurons coding for different visual attributes. Given a large set of geometric information regarding the appearance of an object or class, there may

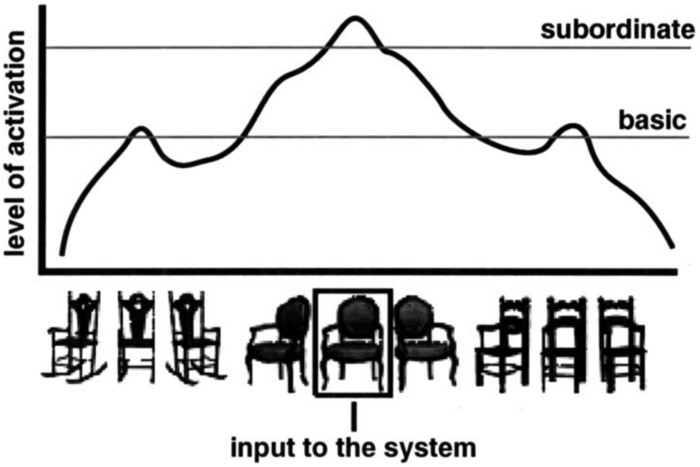

Fig. 9. An exemplar-based mechanism for view and class generalization. Upon presentation of an object, units coding for similar views of the same object and similar views of similar objects are activated. Depending on the specificity of the judgment required, the system can derive either a coarse or a progressively finer match by varying the threshold of pooled activation (see also Edelman, 1995).

exist sufficient complementary information for inferring the virtual appearance an object at novel viewpoints in depth. Thus, views may sometimes be based on the geometric overlap among visible features rather than absolute differences in viewpoint – an interpretation also consistent with the findings of Perrett et al.

5.3. Conclusions

To summarize our results, we find evidence that viewpoint-specific object representations are apparently learned according to three distinct principles:

1. How frequently a given object appears at a given viewpoint.
2. How frequently visually-similar objects appear at a given viewpoint.
3. How dramatically the geometric image structure of objects changes at different viewpoints.

Earlier models of image-based recognition assumed that viewpoint invariance was a consequence of practice with specific viewpoints and that viewpoint-dependent performance at unfamiliar viewpoints was best explained by normalization to familiar viewpoint-specific representations (Bülthoff and Edelman, 1992; Tarr, 1995). The inherent view-specificity of this account seemed to associate, at least implicitly, the mechanisms used for view generalization with inflexible templates, a point reinforced by related computational models of the time (Poggio and Edelman, 1990; Weinshall et al., 1990). Motivated by the well-known limitations of standard template models, proponents of alternative approaches claimed that image-based models were incapable of generalizing across class (Biederman and Gerhardstein, 1993, 1995), a problem at least as challenging as generalizing across viewpoint. To the extent that no specific image-based model proposed a mechanism for achieving class generalization, this criticism remained valid. Recent computational work, however, indicates that this limitation does not apply to more sophisticated image-based approaches (Poggio and Brunelli, 1992; Lando and Edelman, 1995; Vetter et al., 1995; Beymer and Poggio, 1996; Edelman et al., 1996). Here we demonstrate that this limitation does not apply to viewpoint-dependent recognition mechanisms in humans either. We found that viewpoint-specific information learned for some members of a homogeneous class generalized to other members of that class. Such results indicate that shape representations in humans are indeed viewpoint-specific, depicting the appearance of an object from distinct views, but that these representations are flexible enough to support a range of recognition tasks, including both fine exemplar-specific discriminations and coarse categorical judgments.

Acknowledgements

The authors wish to thank Laurie Heller, Shimon Edelman, and Tommy Poggio for their comments, William Milberg and Sheila Blumstein for their help in interpreting the data, Scott Yu for creating the objects used in Experiment 3, and Carmita

Signes and Tom Andrus for help in running some of the experiments. This work was supported by NSF grant BNS 8518774 to Steven Pinker, MIT, and grants from the Yale University Social Science Faculty Research Fund and the Air Force Office of Scientific Research, contract number F49620-91-J-0169 to MJT.

References

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

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., Cooper, E.E., 1991. Evidence for complete translational and reflectional invariance in visual object priming. Perception 20, 585–593.

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.

Brown, R., 1958. How shall a thing be called?. Psychological Review 65, 14–21.

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

Bülthoff, H.H., Edelman, S.Y., Tarr, M.J., 1995. How are three-dimensional objects represented in the brain?. Cerebral Cortex 5 (3), 247–260.

Cohen, D., Kubovy, M., 1993. Mental rotation, mental representation, and flat slopes. Cognitive Psychology 25 (3), 351–382.

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

Corballis, M.C., Zbrodoff, N.J., Shetzer, L.I., Butler, P.B., 1978. Decisions about identity and orientation of rotated letters and digits. Memory and Cognition 6, 98–107.

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., Cutzu F., Duvdevani-Bar, S., 1996. Similarity to reference shapes as a basis for shape representation. In: Cottrell, G.W. (Ed.), Proceedings of 18th Annual Conference of the Cognitive Science Society. San Diego, CA, pp. 260–265.

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

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

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

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.

Hayward, W.G., 1998. Effects of outline shape in object recognition. Journal of Experimental Psychology: Human Perception and Performance, 24 (2) 427–440.

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.

Hill, H., Schyns, P.G., Akamatsu, S., 1997. Information and viewpoint dependence in face recognition. Cognition 62 (2), 201–222.

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

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

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., Gluck, M., Kosslyn, S.M., 1984. Pictures and names: making the connection. Cognitive Psychology 16, 243–275.

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.

Koenderink, J.J., 1987. An internal representation for solid shape based on the topological properties of the apparent contour. In: Richards, W., Ullman, S. (Eds.), Image Understanding 1985–86. Norwood, NJ: Ablex Publishing Corporation, pp. 257–285.

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

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.

Librande, S., 1992. Example-based character drawing. Unpublished Master's thesis, School of Architecture and Planning, Massachusetts Institute of Technology, Cambridge, MA.

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.

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

Miyashita, Y., Date, A., Okuno, H., 1993. Configurational encoding of visual forms by single neurons of monkey temporal cortex. Neuropsychologia 31, 1119–1132.

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

Murray, J.E., Jolicoeur, P., McMullen, P.A., Ingleton, M., 1993. Orientation-invariant transfer of training in the identification of rotated natural objects. Memory and Cognition 21 (5), 604–610.

O'Toole, A., Edelman, S., Bülthoff, H.H., 1998. Stimulus-specific effects in face recognition over changes in viewpoint. Vision Research, in press.

Palmer, S., Rosch, E., Chase, P., 1981. Canonical perspective and the perception of objects. In: Long, J., Baddeley, A. (Eds.), Attention and Performance IX. Hillsdale, NJ: Lawrence Erlbaum, pp. 135–151.

Parsons, L.M., 1987. Visual discrimination of abstract mirror-reflected three-dimensional objects at many orientations. Perception and Psychophysics 42 (1), 49–59.

Perrett, D.I., Oram, M.W., Wachsmuth, E., 1998. Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. Cognition, in press.

Poggio, T., Brunelli, R., 1992. A novel approach to graphics (Technical Report No. 1354). Massachusetts Institute of Technology, USA.

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 (Technical Report No. 1347). Massachusetts Institute of Technology, USA.

Rock, I., 1973. Orientation and Form. New York: Academic Press.

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

Schyns, P.G., 1998. Diagnostic recognition: task constraints, object information, and their interactions. Cognition, in press.

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., Zabinski, M., Blanz, V., 1997. To what extent do unique parts influence recognition across changes in viewpoint?. Psychological Science 8 (4), 282–289.

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., 1991. Orientation-dependent mechanisms in shape recognition: further issues. Psychological Science 2 (32), 207–209.

Troje, N., Bülthoff, H.H., 1996. Face recognition under varying pose: The role of texture and shape. Vision Research 36 (12), 1761–1771.

Tversky, B., Hemenway, K., 1984. Objects, parts, and categories. Journal of Experimental Psychology: General 113, 169–193.

Ullman, S., 1998. Three-dimensional object recognition based on the combination of views. Cognition 67, 21–44.

Ullman, S., Basri, R., 1991. Recognition by linear combinations of models. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (10), 992–1006.

Vetter, T., Hurlbert, A.M., Poggio, T., 1995. View-based models of 3D object recognition: invariance to imaging transformations. Cerebral Cortex 2 (3), 261–269.

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.

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

Yin, R.K., 1969. Looking at upside-down faces. Journal of Experimental Psychology 81 (1), 141–145.