

Viewer-external frames of reference in the mental transformation of 3-D objects

FLORIAN WASZAK

Max Planck Institute for Human Cognitive and Brain Sciences, Munich, Germany
Laboratoire de Psychologie Expérimentale, CNRS
and René Descartes University-Paris 5, Boulogne Billancourt, France

KNUT DREWING

University of Giessen, Giessen, Germany

and

RAINER MAUSFELD

University of Kiel, Kiel, Germany

Most models of object recognition and mental rotation are based on the matching of an object's 2-D view with representations of the object stored in memory. They propose that a time-consuming normalization process compensates for any difference in viewpoint between the 2-D percept and the stored representation. Our experiment shows that such normalization is less time consuming when it has to compensate for disorientations around the vertical than around the horizontal axis of rotation. By decoupling the different possible reference frames, we demonstrate that this anisotropy of the normalization process is defined not with respect to the retinal frame of reference, but, rather, according to the gravitational or the visuocontextual frame of reference. Our results suggest that the visual system may call upon both the gravitational vertical and the visuocontext to serve as the frame of reference with respect to which 3-D objects are gauged in internal object transformations.

Object recognition is one of the most impressive abilities of the human organism. The visual system recognizes objects even though the retinal image varies following rotations of the object or following changes of the observer's viewing position. In fact, depending on its orientation, an object can correspond to an infinite number of different retinal projections. This ambiguity between the distal and the proximal stimulus is the elemental problem the visual system has to deal with (not only in object recognition).

Current Theories of Object Recognition

Contemporary theories of object recognition differ in how they consider the visual system solves this ambiguity. "Structural description models" assume that some specific features of the retinal projections of objects are still relatively independent of the object-observer relationship. Accordingly, the identification of these features is sufficient for the identification of the object (for details, see Biederman, 1987; Hummel & Biederman, 1992). These models predict that recognition performance should be invariant regarding spatial transformations. However, a

large number of studies demonstrate that the identification of objects depends on their current orientation. Objects are recognized faster when presented from a canonical perspective ("three-quarters view"; Blanz, Tarr, & Bühlhoff, 1999). Moreover, the more disoriented an object is with respect to the canonical perspective, the more time it takes to identify the object (Bühlhoff, Edelman, & Tarr, 1995; Tarr & Bühlhoff, 1998). For novel objects, new perspectives may become canonical, if subjects are presented repeatedly with these views. Tarr (1995) and Tarr and Pinker (1989) showed subjects novel objects from specific orientations. With increasing practice, subjects were able to recognize the objects equally quickly from all practiced perspectives. However, when the objects were suddenly presented from a new perspective, recognition performance depended on the distance from the nearest practiced view. Such effects of the object-observer relationship have been demonstrated in a variety of situations: for rotations in the picture plane (e.g., Jolicœur, 1988) and rotations in the depth plane (e.g., Lawson & Humphreys, 1998; Lawson, Humphreys, & Jolicœur, 2000); for novel objects (see also Bühlhoff & Edelman, 1992; Edelman & Bühlhoff, 1992), and for familiar objects (e.g., Hayward & Tarr, 1997; Lawson & Humphreys, 1998); on the subordinate level of categorization (e.g., Edelman & Bühlhoff, 1992; Tarr, 1995), and for the entry level of recognition (Hayward & Williams, 2000; Lawson & Humphreys, 1998). All in all, there is, thus, compelling evidence that

This study was funded by a grant from the DFG (the German national science foundation). The authors thank Dirk Sommer for programming the software. Correspondence should be addressed to F. Waszak, Max Planck Institute for Human Cognitive and Brain Sciences, Amalienstr. 33, 80799 Munich, Germany (e-mail: f.waszak@gmx.net).

the recognition of objects systematically depends on the object's orientation: The more the actual perspective of the object departs from a canonical or familiar view, the worse the recognition performance (for some physiological evidence, see, e.g., Logothetis & Pauls, 1995).

Viewpoint-dependent models of object recognition account for these results. They assume that object recognition is based on the matching of the 2-D view of the object with descriptions of the object stored in memory. Tarr and Pinker (1989), for example, proposed that a collection of 2-D views is stored in memory. Furthermore, this theory (and all other theories of this kind) propose an incremental (time-consuming) normalization process that compensates for the differences between the 2-D percept and one of the stored representations. What kind of normalization exactly takes place is a matter of debate. The most "simple" normalization account assumes that an incremental transformation similar to mental rotation establishes the match between percept and memory representation (e.g., Jolicœur, 1990; Tarr, 1995; Tarr & Pinker, 1989). "Mentally rotating" an object means rotating an internal representation of the object in a way that is similar to physically rotating the object (Shepard & Metzler, 1971).

Before we go on, we need to make a caveat about the generalizability of the results to be presented below. In a mental rotation task, subjects typically have to make a "handedness" judgment (i.e., to discriminate between mirror reflections of a 2-D or a 3-D object). By contrast, in a typical object recognition task, subjects have to tell whether or not two objects have the same shape. Direct comparisons of these two tasks using the same stimuli revealed a notable similarity in the rate of "normalization" (e.g., Tarr & Pinker, 1989). As mentioned above, this has led to the assumption that mental rotation processes are integral to object recognition (Jolicœur, 1990; Tarr, 1995; Tarr & Pinker, 1989; see also the linear combinations of views account from Ullman, 1998, and Ullman & Basri, 1991). However, there is recent evidence that the two types of tasks may be based on different neural substrates, despite the very similar pattern of viewpoint-dependent performance (Gauthier et al., 2002; Perrett, Oram, & Ashbridge, 1998). In the same vein, some accounts of object recognition assume that normalization procedures used for object recognition are different from those used for mental rotation (Tarr & Bülthoff, 1998) or do not rely on normalization at all (e.g., Bülthoff & Edelman, 1993; Poggio & Edelman, 1990). As will become clear soon, our paradigm is somewhat different from both the mental rotation and the object recognition task. However, one might argue that it resembles more a mental rotation task than an object recognition task. Therefore, strictly speaking, we cannot prove beyond doubt that our findings tap into both mental rotation and object recognition. Further research involving more typical object recognition tasks needs to be done to clarify this issue. However, given the preliminary nature of the debate about whether or not compensation processes in recognition and mental imagery should be equated, we do not differentiate between the two concepts. But it is important to stress that the results reported

below cannot offhand be generalized to object recognition but may be restricted to imagery tasks.

Reference Frames in Visual Cognition

Reference frames play an important role in a multitude of phenomena of cognitive psychology (Hinton & Parsons, 1981; Humphreys, 1983; Palmer, 1989; Palmer, Rosch, & Chase, 1981; Robertson, Palmer, & Gomez, 1987; for reviews, see Hinton & Parsons, 1988; Palmer, 1999; Wraga, Creem, & Proffitt, 1999). All models of object recognition specify a frame of reference with respect to which the object properties are gauged. Two different types of reference frame have been proposed.

An object-centered reference system represents the object's properties with respect to a frame that is inherent to the object itself—for example, an axis of symmetry (Marr & Nishihara, 1978; Sekuler & Swimmer, 2000; but see Large, McMullen, & Hamm, 2003). Structural description models presume object-centered reference frames. These kinds of representation do not change with alterations of the object–observer relationship, but they are difficult to derive from the retinal input. Certainly, psychophysical experiments have not found much evidence for this kind of view-independent representation. However, there is an ongoing debate as to the conditions under which the visual system may use viewpoint-variant representations during visual recognition (see Wilson & Farah, 2003). By contrast, view-dependent models of object recognition imply a viewer-centered reference system representing object properties in relation to the observer, most often in retinotopic coordinates. That is, the object is represented in "views," which correspond to the object's retinal projection. Evidently, viewer-centered representations are easily derived from the proximal stimulus, but each object–observer relationship yields a different representation. As described above, normalization processes are assumed to deal with this problem.

Research has focused on the question as to whether reference frames are viewer centered or object centered. However, research in other domains of visual perception has accumulated a lot of evidence that viewer-external frames of reference other than the object-centered frame play an important role in visual perception. One example is the reference frame of the visual context (for reviews, see Hinton & Parsons, 1988; Palmer, 1999). Objects in the real world are almost always accompanied by other objects forming a global context or scene (Biederman, 1972). Visual contexts contain complex structures of covariation between visual objects, in which regularities exist and which are, by and large, stable over time. A room, for example, has a door and a window in the wall, a table and chairs on the floor, and a lamp hanging from the ceiling. Another example is that of the gravitational vertical. Evidently, these meaningful invariants in the visual world contain very useful information, which the visual system might harness for the identification of objects.

However, in a typical object recognition paradigm, subjects are seated upright in front of the monitor with the visual context not being controlled for (or being completely

masked). Evidently, under such conditions, the retinal, the gravitational, and the visuocontextual reference frames are utterly confounded. To investigate with respect to which of these possible frames objects are actually represented, we would need to decouple the frames by varying the position of the observer and/or the orientation of the visual context.

Just a few studies have done so, most of which date back to the 1970s, the 1960s, and even the 1950s, and all of which have been limited to 2-D figures/rotations in the picture plane. Rock and Heimer (1957) presented subjects with novel figures in a study and a recognition phase. In the study phase, each subject sat upright; in the recognition phase, the subject's head was tilted by 90°, such that the retinotopic and the gravitational/visuocontextual reference frames were decoupled. Subjects' recognition performance was better when the figures were presented upright (i.e., oriented as in the study phase) with respect to the gravitational/visuocontextual vertical than when the figures were presented upright in retinotopic coordinates. Similarly, Rock (1956) presented subjects whose heads were tilted by 90° with "ambiguous figures" that depicted two different objects, depending on whether they were perceived with respect to the gravitational/visuocontextual reference (e.g., a dog) or with respect to the retinotopic reference (e.g., a cook). Subjects interpreted the figures in more than 80% of all trials with respect to the gravitational/visuocontextual vertical. This was also the case when the visuocontext was completely masked. Corballis, Zbrodoff, and Roldan (1976) showed that subjects whose heads were tilted by 60° rotated alphanumeric characters into the gravitational upright position when asked to make a handedness discrimination. Thus, *upright* was defined by the gravitational rather than by the retinal vertical. McMullen and Jolicœur (1990) reported that subjects tend to rotate into the retinal upright when naming line drawings of objects and into the environmental upright (gravitation plus aligned border) when doing left-right reflection discriminations. Taken together, these studies suggest that both the gravitational and the visuocontextual frames of reference play a role in 2-D object recognition. A number of experiments have demonstrated visual representations in retinotopic coordinates (see Banks & Stolarz, 1975; Corballis & Roldan, 1975; Gillam & McGrath, 1979; Maffei & Campbell, 1970; Wade, 1973). However, these experiments dealt with low-level perception, such as afterimages, judgments of symmetry, and visual acuity. In a review of the literature, Hock and Sullivan (1981) suggested that visual objects are not necessarily represented in gravitational coordinates, but they may be so (1) if the task involves information to be stored in memory or to be retrieved from memory, (2) if the stimulus material is not too complex, and (3) if neither the instruction nor the visual context makes subjects use a retinal frame of reference.

Hock and Sullivan's (1981) third condition was based on experiments demonstrating that, for a handedness discrimination, subjects mentally rotate letter-like figures into the retinally upright position if the canonical vertical of a rectangle (i.e., the visuocontextual frame of refer-

ence) is aligned with the subjects' retinal vertical (Corballis, Nagourney, Shetzer, & Stefanatos, 1978; see also Asch & Witkin's, 1948, "rod and frame effect").

Anisotropy in Normalization Processes

Most previous studies have looked at the frame of reference for rotations around the z-axis (i.e., for picture plane rotations). In most of these studies, subjects were presented with familiar 2-D objects with distinct "tops" (i.e., with an intrinsic axis of orientation), while their retinal frame of reference and the gravitational/visual frames of reference were decoupled. The reasoning in all these studies is as follows (although the procedures might be different). Recognition performance should be better when the object-intrinsic axis of orientation aligns with the frame of reference used by the subjects to represent the object. Now, if recognition reaction times (RTs) are faster when the object-intrinsic axis aligns with, say, the gravitational/visual frame than when it aligns with the retinal frame, then it can be inferred that the subjects use the gravitational/visual frame of reference. This type of reasoning is perform bound to 2-D object recognition and rotations in the picture plane.

In the present study, by contrast, we tried to examine the frame of reference with a completely different experimental procedure that did not rely on 2-D object recognition and picture plane rotations. Instead, our experimental procedure was based on 3-D recognition and depth rotations. To do so, we took advantage of an anisotropy of recognition performance: As described above, the recognition performance depends on the *extent* of misalignment between a test and a canonical or familiar view of an object. However, a number of studies suggest that it additionally depends on the *direction* of misalignment. Bühlhoff and Edelman (1992) found smaller error rates for disorientations around the y-axis (vertical) relative to disorientations around the x-axis (horizontal). Similarly, Parsons (1987) reported that mental rotations around the y-axis were faster than those around the x-axis.

This anisotropy of recognition performance is the key to examining the frame of reference in 3-D recognition. We tested the subjects' object recognition performance for disorientations around the y-axis and the x-axis, while the retinal frame of reference and the gravitational frame of reference were decoupled (i.e., when the observer was tilted by 90°). If, so the reasoning goes, the pattern of results (y-axis faster than x-axis) does not change relative to when the retinal and the gravitational frames of reference are aligned (i.e., when the observer sits upright in front of the screen), then the gravitational vertical serves as the frame of reference with respect to which objects are gauged in 3-D object recognition. If, in contrast, the pattern of results switches (x-axis faster than y-axis) when the observer is tilted, then the subjects use the retinal frame of reference. As far as we know, this novel approach allows the frame-of-reference question in 3-D recognition to be explored for the first time.

Another important aspect of our study concerns the assessment of the anisotropy of recognition performance.

All experiments outlined above confounded the axis of disorientation and the proximal stimulation presented to the observer. In these experiments, subjects were simply presented with a view of an object in a standard orientation and test views of the same object misaligned by rotations around the different axes. Evidently, differences in recognition performance might have been due not to the axis of rotation but rather to differences in the stimulus material (i.e., the test views), the more so as the different views of the objects were not equally complex. However, we conceived a new paradigm that avoided this shortcoming (see below).

EXPERIMENT

The present study belongs to the tradition of research that uses novel, "meaningless" objects (i.e., objects that the observer has never seen before; e.g., Bühlhoff & Edelman, 1992; Bühlhoff et al., 1995; Hayward & Tarr, 1997; Jolicœur & Humphrey, 1998; Poggio & Edelman, 1990; Tarr, 1995; Tarr & Bühlhoff, 1998; Tarr & Pinker, 1989; Ullman, 1989; Ullman & Basri, 1991; see also Cooper, 1976; Shepard & Metzler, 1971). In this line of research, meaningless objects have been used because researchers wanted to investigate object recognition "in its pure form" without having to contend with memory effects. Basically, the main argument for taking novel, meaningless objects is that, for familiar objects, an uncountable number of views from all possible perspectives is already stored in memory, such that possible effects of the (viewer-centered) perspective are washed out. The results of this kind of study are the basis for all viewer-centered theories of object recognition.

The aim of the present study was to investigate whether the "representational space" in which normalization processes take place is defined in retinotopic coordinates, as assumed by most (if not all) view-dependent theories of object recognition or in gravitational coordinates or coordinates of the visual context. We investigated subjects' recognition performance for disorientations around different axes of rotation (vertical or horizontal). At the same time, we varied the mutual orientation of the retinal, the gravitational, and the visuocontextual reference frames. Following the literature on 2-D recognition, we expected that—with the visuocontext being masked—the representational space is gauged in gravitational coordinates (i.e., that subjects respond faster when an object is misaligned around the gravitational vertical axis relative to when an object is misaligned around the gravitational horizontal axis, irrespective of whether or not the gravitational and the retinal vertical/upright are coupled). However, we also expected that a visuocontext may be able to overcome the gravitational reference frame—that is, if a visuocontext is present, subjects will respond faster when an object is misaligned around the canonical vertical axis of the visuocontext (cf. findings from Corballis et al., 1978).

One trial of the experiment comprised the following sequence of events: First, the subjects were presented with

a first view ("frontal view") of an object. Next, a cue indicated from which perspective the object might be presented next ("target view"). Possible target views were the frontal view rotated by 90° or -90° around the x - or y -axis (see Figure 1). The cue was also intended to mask the visual short-term memory (Sperling, 1960). Subsequently, the second (test) view was presented; the subjects had to decide as fast as possible whether or not it corresponded to the target view.

This procedure differs slightly from usual paradigms in object recognition research (for a similar procedure, see Shepard & Cooper, 1982). The most common paradigm is probably the sequential matching task. In this task, subjects see in quick succession two objects, which are presented in different viewpoints. The subjects' task is simply to decide whether the two objects are the same or different, regardless of any rotation. However, in the present study, we could not draw on this procedure in its standard form, because any difference in recognition performance between the two axes of rotation (x and y) revealed by this procedure cannot be interpreted unambiguously. The reason for this is that the normalization process possibly tries by default at first to find a match for differences in viewpoint that are due to a y -axis rotation and only thereafter, if this turned out to be unsuccessful, for differences in viewpoint that are due to an x -axis rotation. Our modified procedure avoided this problem, in that the axis of rotation around which the two views were misaligned was cued in advance. This guaranteed that any difference in recognition performance between the two axes of rotation (x and y) was not confounded by a possible fixed order in which the normalization process takes effect.

There were three different experimental conditions. In Condition 1, the subjects sat upright in front of the monitor, as in usual object recognition experiments. In Condition 2, they were tilted by 90° . In Conditions 1 and 2, the visuocontext was completely masked. In Condition 3, the subjects were tilted by 90° , as in Condition 2. However, in this condition, the objects were surrounded by a visual context (see Figure 3), which also was tilted by 90° (i.e., the context's canonical vertical was aligned with the subject's [tilted] retina). In Conditions 2 and 3, the response pad was tilted as well. Figure 1 shows the three experimental conditions.

Method

Subjects. Two of the authors (F.W. and K.D.) and 3 naive subjects volunteered to participate in the experiment. All subjects had normal or corrected-to-normal vision.

Apparatus. The experiment was carried out on an IBM-compatible PC (Intel CPU 686-200 MHz, Elsa GL 8 MB). Stimuli were displayed on a 21-in. SVGA computer screen (Eizo F764-T92) with $1,024 \times 768$ pixel spatial resolution and 75-Hz temporal resolution. The subjects responded by pressing keys on a game pad. The experiment was controlled by custom-made software.

The monitor was fit into a "box," at the end of which was affixed a diving mask. The viewing distance was 130 cm. The visual angle of the objects' projection on the subjects' retinæ was about $6^\circ 60' \times 6^\circ 60'$. The inside of the box was lined with black fabric. The subjects looked through the diving mask on the screen. This apparatus

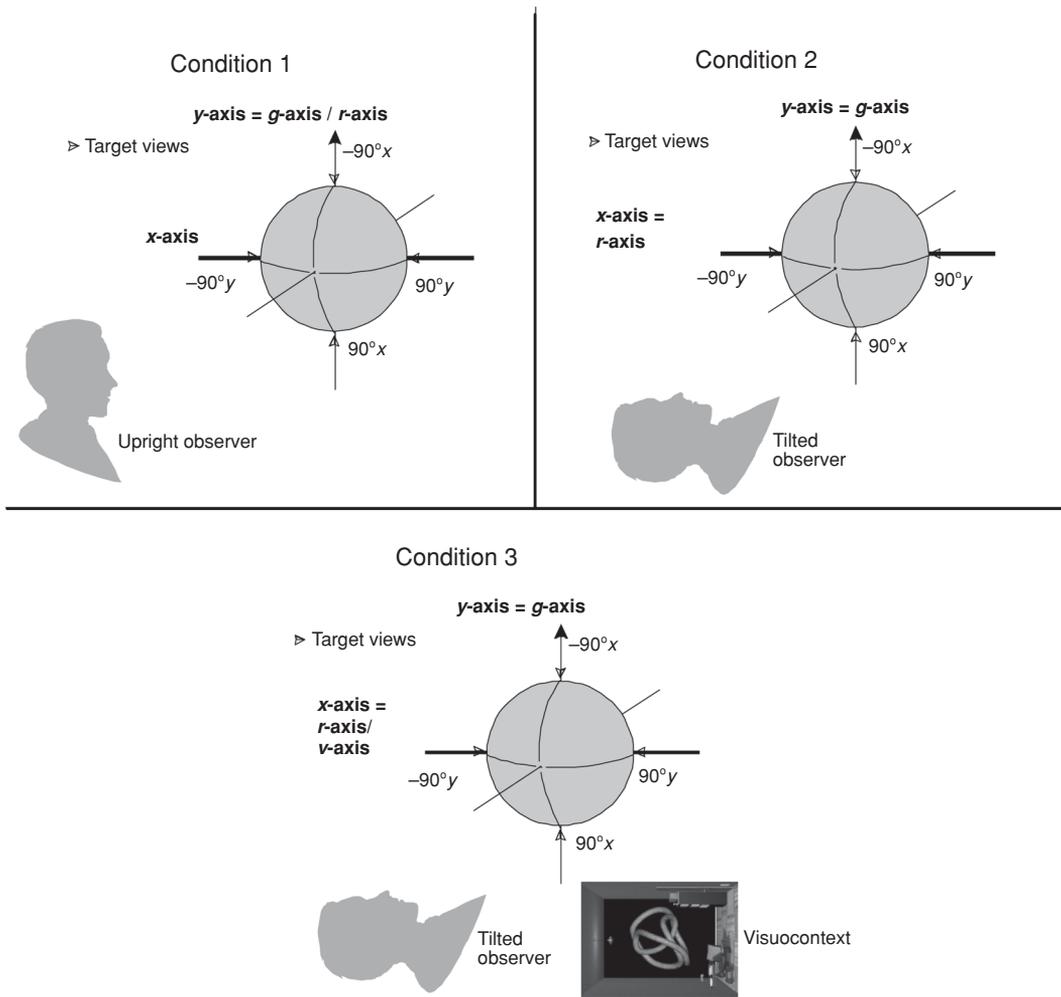


Figure 1. Definition of the axes' labels and visualization of the decoupling of the reference frames. Independent from the experimental condition, the axis vertical with respect to the upright observer is labeled “ y -axis.” The axis perpendicular to the y -axis and to the line of sight is labeled “ x -axis.” In all conditions, the observer was successively presented with a frontal view, a cue indicating the target perspective, and a test view differing from the frontal view by a rotation of 90° or -90° around the x - or y -axis. The subject’s task was to decide whether or not the test view corresponded to the target view. In Condition 1, the observer sat upright in front of the monitor. In Condition 2, the observer was tilted by 90° . In Conditions 1 and 2, the visual context was completely masked. In Condition 3, the observer was tilted by 90° , and a visual context was displayed aligned with the x -axis. The r -, g -, and v -axes denote the retinal, the gravitational, and the visuocontextual vertical, respectively.

ensured that the subjects saw nothing but the stimuli presented on the screen, making the experimental control of the visuocontext possible.

Stimuli. Twelve objects were constructed with the aid of the software package 3D Studio Max (Autodesk, Inc., 2000). Each object’s surface was yellow-brown and textured. They were presented against a black background. All objects had a similar global shape (“tube-like”). For each object, one view was selected that served as the frontal view. This frontal view was chosen such that all parts of the objects were visible. Apart from that, the view was arbitrarily assigned. In order to improve the perception of 3-D structure, the frontal views were slightly animated. The motion sequence consisted of 20 views of the objects, all of which resulted from a 2° rotation of the frontal view around 1 of 20 axes parallel to the picture plane and intersecting at the center of gravity of an object. The position of each of the axes in the picture plane differed by 18° ($360^\circ/20 = 18^\circ$). That

is, if the first view of the motion sequence resulted from a 2° rotation around, say, the x -axis, then the second view resulted from a 2° rotation around the axis that differed from the x -axis by a rotation of 18° in the picture plane, and so on. This sequence warranted that the motion did not have any particular direction in space. On the basis of the frontal view, another four (side) views were rendered that were defined by rotating the object by 90° or -90° around the x - or y -axis, respectively (see Figure 1). That is, for each object, four view pairs were rendered, each pair comprising the frontal view and one of the side views. Figure 2 demonstrates that by rotating the two views of a pair in the picture plane, we were able to use each pair for each of the four possible target views. By doing so, we avoided a situation in which the axis of disorientation and the proximal stimulation are confounded: In contrast to the experiments mentioned above, we presented the same pairs of views for all axes of rotation.

Visuocontext. The visuocontext presented on the screen in Condition 3 is shown in Figure 3. It was a virtual room that was furnished with a couple of objects, all of which had a canonical vertical. The experimental objects were displayed against the black back wall (see Figure 3). The room was rendered with $1,024 \times 768$ pixel spatial resolution. We used this virtual miniature room rather than a simple rectangle with a canonical vertical because we considered a visuocontext with unambiguous top-bottom relations familiar from everyday life to be more likely to be called upon to serve as a frame of reference.

Procedure. According to the experimental condition, the subjects either sat upright (Condition 1) or lay tilted by 90° in front of the monitor (Conditions 2 and 3). The subjects initiated the self-paced trials by pressing a key on the game pad. Then, the frontal view was presented for 5,000 msec. Immediately thereafter, the cue indicating the target view was presented for 250 msec (see Figure 4). Finally, the test view was presented until the subject responded (or until 2,000 msec had elapsed). The test view showed the same object as the frontal view either in the target view or in the distractor view. The distractor view corresponded to the view opposite to the target view (i.e., if the target view was defined by the rotation $90^\circ y$, the distractor view was defined by the rotation $-90^\circ y$). The subjects' task was to decide as fast as possible whether the test view showed the object in the target view or in the distractor view.

Design. In each of the three conditions, 6 of the 12 objects (i.e., the four view pairs of the objects) were used. The objects were counterbalanced across the conditions such that each condition shared 2 objects with each of the other two conditions (i.e., 4 objects were shared and the remaining 2 were presented exclusively in the given condition). In each condition, each view pair was presented eight times, once for each of the four target views and once for each of the four distractor views. This design ensured that the proximal stimulation was the same for all target (and distractor) views and, hence, that any effect between the different target views could be attributed to the absolute difference of misalignment between frontal and test view (x -axis vs. y -axis). Overall, each condition comprised 6 (objects) \times 4 (view pairs) \times 8 (= 4 target + 4 distractor views) = 192 trials.

The 192 trials of each condition were performed in two blocks. The six experimental blocks were performed in three different sessions. Each session lasted about 1.5 h. The blocks were counterbal-

anced across sessions. Within each block, the sequence of trials was randomized.

Results and Discussion

Incorrect responses and RTs more than 1.5 *SD* above the grand mean (6.9% of the data) were eliminated before the data analysis. Figure 5 shows RTs in the three experimental conditions separately for the two axes of rotation (x -axis, y -axis). Table 1 shows the individual RT means for each condition \times axis combination. An ANOVA was run on the error data, including the factors condition (1–3) and axis of rotation (x -axis vs. y -axis). Although the ANOVA yield a significant main effect of axis of rotation [$F(1,4) = 12.02, p < .05$], t tests did not reveal differences between the axes of rotation to be significant in any of the conditions (Condition 1, $y = 11\%$, $x = 12\%$; Condition 2, $y = 11\%$, $x = 13\%$; Condition 3, $y = 19\%$, $x = 19\%$). Moreover, the error pattern did not counteract the RT pattern. Thus, a speed-accuracy trade-off within the conditions can be excluded. In Condition 3, the subjects committed more errors and reacted a bit faster than in the other two conditions. However, notice that any possible speed-accuracy trade-off across conditions did not affect the main results of the experiment, which relied on the pattern of results within the conditions.

An ANOVA including the factors condition (1–3) and axis of rotation (x -axis vs. y -axis) was also run on the RT data. The only significant effect was the interaction of condition and axis of rotation [$F(2,8) = 7.01, p < .02$]. One-tailed t tests showed the difference between the axes to be significant for Conditions 1 and 2 and almost significant for Condition 3 [$t(4) = 3.15, p < .02$; $t(4) = 2.99, p < .02$; $t(4) = 1.97, p < .06$, respectively].

First of all, the results show that the subjects' performances depended on the direction of disorientation between the frontal view and the test view. In all three

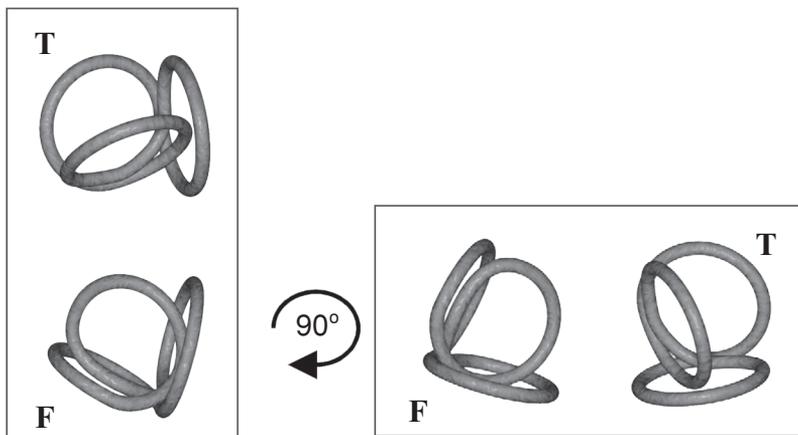


Figure 2. The figure shows how the proximal stimulation was controlled for across the different disorientations. By rotating the views in the picture plane, the same pair of views can be presented for all four disorientations ($-90^\circ/90^\circ, y$ -axis/ x -axis). The figure shows the same pair of views used for the $-90^\circ x$ and $90^\circ y$ disorientations, respectively. F denotes the frontal views; T denotes the target views.

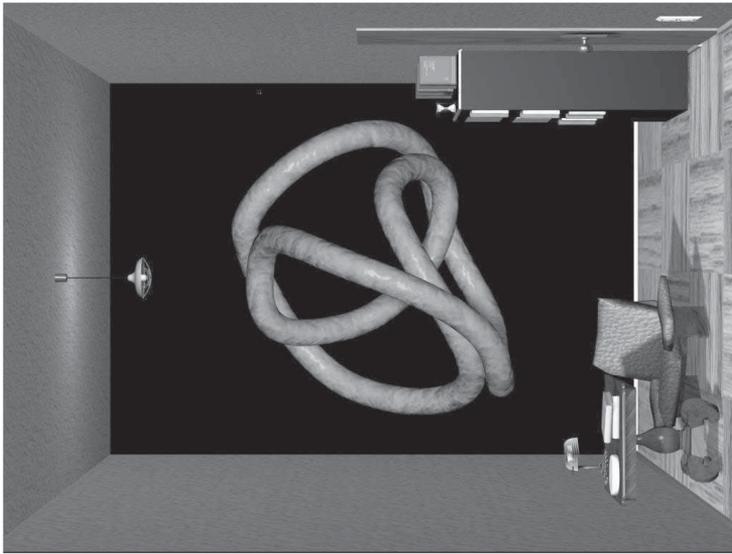


Figure 3. The visuocontext presented in Condition 3.

conditions, there was an RT difference between disorientations around the x -axis and disorientations around the y -axis. Note that this confirms an anisotropy of the representational space without the direction of misalignment and the proximal stimulation being confounded. This is in line with results from previous studies using the same experimental paradigm (Drewing, Waszak, & Mausfeld, 1997), which revealed similar direction dependencies in RTs when one is the test view resulted from the frontal view by depth rotations of less than 90° (30° , 50° , and 70°). The latter finding demonstrates that the anisotropy relates to the axis of a normalization process rather than to a certain target perspective. Taken together, these results highlight that, when modeling normalization processes, it is not sufficient to take into account the relative difference of disorientation (i.e., the degree of misalignment), as it is done by most models relying on normalization processes. Such theories predict a dependency of recognition performance on the disorientation between the percept and the object representation stored in memory (e.g., Tarr & Pinker, 1989; Ullman, 1989; Ullman & Basri, 1991), but they do not usually predict an influence of the direction of disorientation (for an exception and computational implementation, see Bülthoff & Edelman, 1992). These theories are based on the idea that normalization is an algorithmic process that matches, in a mathematically optimizing manner, the image-like information of the perceptual and the memory representation. However, if view dependency was exhaustively described by the mere difference between image-like information carried by the representations, there should be no performance difference between the different axes of rotation. The present results show this idea to be oversimplified.

The anisotropy of performance demonstrated in our experiment may be related to a series of experiments from Proffitt and colleagues (Carpenter & Proffitt, 2001;

Wraga, Creem, & Proffitt, 2000). Wraga et al. (2000) showed that subjects are better at imagining a rotation of their own viewpoint than a rotation of the object itself when they are asked to judge what an array of objects in the world looks like from another perspective. Carpenter and Proffitt (2001) tested rotations in alternative planes and found that the viewer rotation advantage occurred only in the transverse (ground) plane, perpendicular to the viewer's principal axis (y -axis). However, viewer rotation performance declined to the level of object rotation performance in the other planes. That is, the results of Carpenter and Proffitt's study also suggest an advantage for rotation around the vertical (defined by gravity and the



Figure 4. The figure shows the cue used to indicate the target view. The cue had approximately the same visual angle as the objects. The cue in the figure indicates the target view $-90^\circ x$ ("from above" with respect to the upright observer).

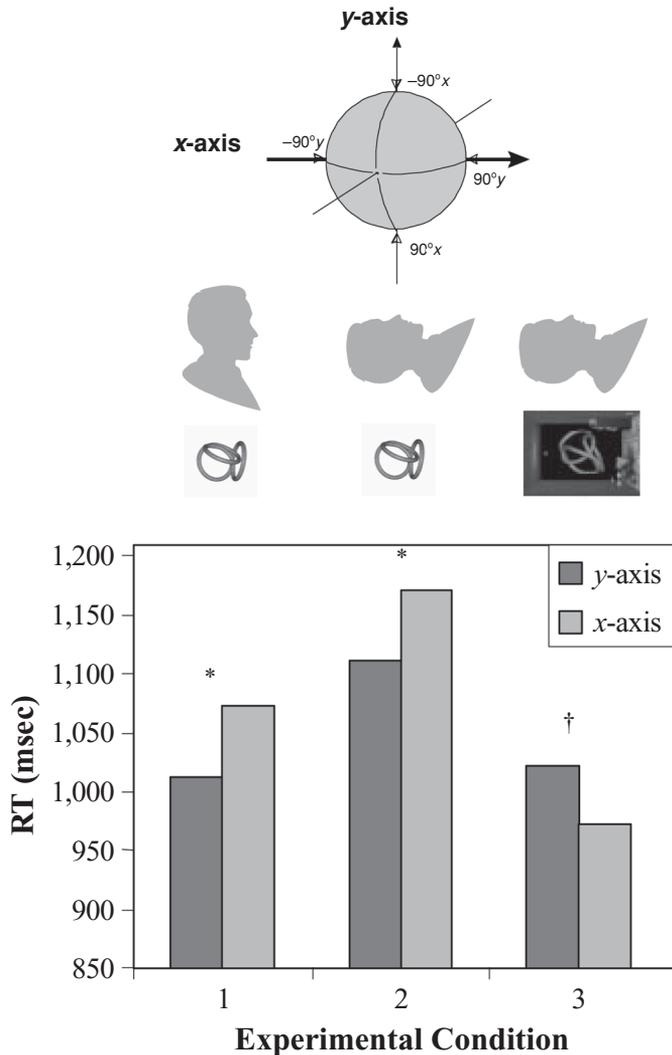


Figure 5. Results. The figure shows reaction times (RTs) separately for the three experimental conditions and the two axes of rotation (x-axis and y-axis). * $p < .05$. † $p < .1$.

visual context). Creem, Wraga, and Proffitt (2001) have suggested that viewer rotations follow geometrical constraints that are ultimately guided by our experience with a gravitational field. Evidently, the normalization procedures examined in the present study might have been influenced by the same constraints, such that, for example, there was an advantage for matching objects with rotated versions that maintained the same “top” of the object.

However, the anisotropy of the normalization performance was primarily meant to help to reveal the frame of reference with respect to which objects are represented. Note that it is impossible to address this question without some feature of the representation indexing distinct directions in space. To our knowledge, this is the first time that this question has been investigated in 3-D normalization.

When the gravitational and the retinal frames of reference were aligned and no visuocontext was present

(Condition 1), RTs were faster for disorientations around the y-axis. That is, views “from the side” were easier to recognize than views “from above/underneath.” Importantly, as expected, this pattern of results did not change when the subjects were tilted by 90° (Condition 2), and, consequently, the retinotopic frame of reference was perpendicular to the gravitational frame of reference. In other words, side views with respect to the gravitational vertical were easier to recognize, even though these views were top–bottom views with respect to retinotopic coordinates.

The pattern of results was reversed when the visuocontext was displayed aligned with the retinal frame of reference of the tilted observer (Condition 3). That is, when aligned with the retinal reference frame, the visuocontext was capable of capturing the supremacy of reference, resulting in faster RTs for disorientations around the x-axis (views from the side with respect to the visuocontext).

Table 1
Mean Reaction Times (RTs; in Milliseconds) for Individual
Subjects Separately for the Three Viewing Conditions and the
Two Axes of Rotation (*y* and *x*)

	Subject	RT	
		<i>y</i>	<i>x</i>
Condition 1	1	923	929
	2	794	905
	3	1,066	1,094
	4	1,278	1,366
	5	999	1,067
Condition 2	1	990	999
	2	1,044	1,110
	3	1,098	1,157
	4	1,233	1,366
	5	1,187	1,228
Condition 3	1	983	968
	2	824	716
	3	955	985
	4	1,494	1,420
	5	855	770

These results suggest that the visual system may call upon both the gravitational vertical and the visuocontext to serve as the frame of reference according to which objects are gauged in 3-D normalization. Certainly, it is convenient to assume that normalization processes rely on retinal coordinates. This assumption makes possible a straightforward investigation of processes that derive information from the proximal stimulation. However, this picture, too, is oversimplified. We suggest that there are several possible reference frames, all of which may take part in representing objects. These frames may be in a kind of hierarchy or be used in some combination. It may be that when two or more reference frames are correlated, this combination of frames is used (as in Condition 3). It may also be that, of all available references, the system selects the one that is most beneficial to the interaction of the observer with the environment. Note that the manipulation of objects (e.g., grasping movements) and locomotion in an environment full of obstacles becomes increasingly easier, the less frequently information represented in the visual system has to be updated. A strictly retinotopic coordination has to be refreshed after any movement of the observer, even a simple movement of the head. Representing objects in reference to the visuocontext, by contrast, allows for larger alterations of the object–observer relationship without necessitating the representation of the environment to be reconstructed time and again. Moreover, data relating to the position of objects in the surrounding that do not depend on the observer's position are the basis for the reliable anticipation of movements of objects.

Interesting in this context is a series of experiments about self-rotations by Creem et al. (2001), who further explored the finding from Carpenter and Proffitt (2001). Carpenter and Proffitt found that the advantage during

imagined viewer rotation (relative to object rotation) holds true only for rotations performed in the horizontal plane (*y*-axis), but not for the other axes of rotation. Creem et al. investigated whether this was a result of physically possible movement obeying gravity or specific to the geometrical relation between the observer and objects. They found a viewer advantage in all cases in which an orthogonal relationship between the viewer and objects was preserved, allowing for a rotation around the viewer's principal axis, regardless of the physical possibility of movement. Hence, these transformations seem to follow principles of geometry and not of physics. However, although gravitational constraints alone did not predict the ease of imagined self-rotations, the results of Creem et al.'s experiments suggest that an advantage for egocentric rotation relies on representing body–environment relations as they are in a physical world constrained by gravity (i.e., egocentric rotations still follow geometrical constraints that are ultimately guided by our experience with a gravitational field).

In the same vein, the present results show that normalization processes cannot be studied in isolation (viz. by focusing on an object's effect on the retina). Rather than considering the main purpose of the normalization processes to consist of the deduction of veridical information about objects, these processes have to be conceived as assisting the process of seeing one's way in the world. To accomplish this task, the visual system not only has to convey information about the identity (object recognition) or handedness (mental rotation) of objects but also has to keep track of their arrangement despite continuous body movements. Representing objects with respect to an external frame of reference may be one of the visual system's means to do so. However, note that an object-centered frame of reference is also not suitable for this purpose. Representing objects with reference to a frame that is located on the observed object makes the representation independent from the mutual orientation of the object and the viewer. However, object-centered representations do not incorporate any information about the whereabouts of the objects in space (nor of the observer).

From this perspective, it is not surprising that the visual system makes use of representations between the extrema of retinotopic representations on one hand, which suffer from the maximum view dependency, and object-centered representations on the other hand, which suffer—so to speak—from a maximum view independency.

REFERENCES

- ASCH, S. E., & WITKIN, H. A. (1948). Studies in space orientation: II. Perception of the upright with displaced visual fields and with body tilted. *Journal of Experimental Psychology*, **38**, 455-477.
- AUTODESK, INC. (2000). 3D Studio Max [Computer software]. San Rafael, CA: Author.
- BANKS, M. S., & STOLARZ, S. J. (1975). The effect of head tilt on meridional differences in acuity: Implications for orientation constancy. *Perception & Psychophysics*, **17**, 17-22.

- BIEDERMAN, I. (1972). Perceiving real-world scenes. *Science*, **177**, 77-80.
- BIEDERMAN, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, **94**, 115-147.
- BLANZ, V., TARR, M. J., & BÜLTHOFF, H. H. (1999). What object attributes determine canonical views? *Perception*, **28**, 575-599.
- BÜLTHOFF, H. H., & EDELMAN, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, **89**, 60-64.
- BÜLTHOFF, H. H., & EDELMAN, S. (1993). Evaluating object recognition theories by computer graphics psychophysics. In T. A. Poggio & D. A. Glaser (Eds.), *Exploring brain functions: Models in neuroscience* (pp. 139-164). New York: Wiley.
- BÜLTHOFF, H. H., EDELMAN, S., & TARR, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, **3**, 247-260.
- CARPENTER, M., & PROFFITT, D. R. (2001). Comparing viewer and array mental rotations in different planes. *Memory & Cognition*, **29**, 441-448.
- COOPER, L. A. (1976). Demonstration of a mental analog of an external rotation. *Perception & Psychophysics*, **19**, 296-302.
- CORBALLIS, M. C., NAGOURNEY, B. A., SHETZER, L. I., & STEFANATOS, G. (1978). Mental rotation under head tilt: Factors influencing the location of the subjective reference frame. *Perception & Psychophysics*, **24**, 263-273.
- CORBALLIS, M. C., & ROLDAN, C. E. (1975). Detection of symmetry as a function of angular orientation. *Journal of Experimental Psychology: Human Perception & Performance*, **1**, 221-230.
- CORBALLIS, M. C., ZBRODOFF, J., & ROLDAN, C. E. (1976). What's up in mental rotation? *Perception & Psychophysics*, **19**, 525-530.
- CREEM, S. H., WRAGA, M., & PROFFITT, D. R. (2001). Imagining physically impossible self-rotations: Geometry is more important than gravity. *Cognition*, **81**, 41-64.
- DREWING, K., WASZAK, F., & MAUSFELD, R. (1997). Untersuchungen zu Struktur und Referenz richtungsabhängiger Effekte perspektivenneutralisierender Operationen. In W. Krause, U. Kotkamp, & R. Goertz (Eds.), *Proceedings der 3. Fachtagung der Gesellschaft für Kognitionswissenschaften* (pp. 25-26). Jena: Friedrich-Schiller-Universität.
- EDELMAN, S., & BÜLTHOFF, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, **32**, 2385-2400.
- GAUTHIER, I., HAYWARD, W. G., TARR, M. J., ANDERSON, A. W., SKUDLARSKI, P., & GORE, J. C. (2002). BOLD activity during mental rotation and viewpoint-dependent object recognition. *Neuron*, **34**, 161-171.
- GILLAM, B., & McGRATH, D. (1979). Orientation relative to the retina determines perceptual organization. *Perception & Psychophysics*, **26**, 177-181.
- HAYWARD, W. G., & TARR, M. J. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **23**, 1511-1521.
- HAYWARD, W. G., & WILLIAMS, P. (2000). Viewpoint dependence and object discriminability. *Psychological Science*, **11**, 7-12.
- HINTON, G. E., & PARSONS, L. M. (1981). Frames of reference and mental imagery. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 261-277). Hillsdale, NJ: Erlbaum.
- HINTON, G. E., & PARSONS, L. M. (1988). Scene-based and viewer-centered representations for comparing shapes. *Cognition*, **30**, 1-35.
- HOCK, H. S., & SULLIVAN, M. (1981). Alternative spatial reference systems: Intentional vs. incidental learning. *Perception & Psychophysics*, **29**, 467-474.
- HUMMEL, J. E., & BIEDERMAN, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, **99**, 480-517.
- HUMPHREYS, G. W. (1983). Reference frames and shape perception. *Cognitive Psychology*, **15**, 151-196.
- JOLICŒUR, P. (1988). Mental rotation and the identification of disoriented objects. *Canadian Journal of Psychology*, **42**, 461-478.
- JOLICŒUR, P. (1990). Identification of disoriented objects: A dual-systems theory. *Mind & Language*, **5**, 387-410.
- JOLICŒUR, P., & HUMPHREY, G. K. (1998). Perception of rotated two-dimensional and three-dimensional objects and visual shapes. In V. Walsh & J. Kulikowski (Eds.), *Perceptual constancy: Why things look as they do* (pp. 69-123). Cambridge: Cambridge University Press.
- LARGE, M.-E., McMULLEN, P. A., & HAMM, J. P. (2003). The role of axes of elongation and symmetry in rotated object naming. *Perception & Psychophysics*, **65**, 1-19.
- LAWSON, R., & HUMPHREYS, G. W. (1998). View-specific effects of depth rotation and foreshortening on the initial recognition and priming of familiar objects. *Perception & Psychophysics*, **60**, 1052-1066.
- LAWSON, R., HUMPHREYS, G. W., & JOLICŒUR, P. (2000). The combined effects of plane disorientation and foreshortening on picture naming: One manipulation or two? *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 568-581.
- LOGOTHETIS, N. K., & PAULS, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, **3**, 270-288.
- MAFFEI, L., & CAMPBELL, F. W. (1970). Neurophysiological localization of the vertical and horizontal visual coordinates in man. *Science*, **167**, 386-387.
- MARR, D., & NISHIHARA, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London: Series B*, **200**, 269-291.
- McMULLEN, P. A., & JOLICŒUR, P. (1990). The spatial frame of reference in object naming and discrimination of left-right reflections. *Memory & Cognition*, **18**, 99-115.
- PALMER, S. E. (1989). Reference frames in the perception of shape and orientation. In B. E. Shepp & S. Ballesteros (Eds.), *Object perception: Structure and process* (pp. 121-163). Hillsdale, NJ: Erlbaum.
- PALMER, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.
- PALMER, S. E., ROSCH, E., & CHASE, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 135-151). Hillsdale, NJ: Erlbaum.
- PARSONS, L. M. (1987). Visual discrimination of abstract mirror-reflected three-dimensional objects at many orientations. *Perception & Psychophysics*, **42**, 49-59.
- PERRETT, D. I., ORAM, M. W., & ASHBRIDGE, E. (1998). Evidence accumulation in cell populations responsive to faces: An account of generalization of recognition without mental transformations. *Cognition*, **67**, 111-145.
- POGGIO, T., & EDELMAN, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, **343**, 263-266.
- ROBERTSON, L. C., PALMER, S. E., & GOMEZ, L. M. (1987). Reference frames in mental rotation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 368-379.
- ROCK, I. (1956). The orientation of forms on the retina and in the environment. *American Journal of Psychology*, **69**, 513-528.
- ROCK, I., & HEIMER, W. (1957). The effect of retinal and phenomenal orientation on the perception of form. *American Journal of Psychology*, **70**, 493-511.
- SEKULER, A. B., & SWIMMER, M. B. (2000). Interactions between symmetry and elongation in determining reference frames for object perception. *Canadian Journal of Experimental Psychology*, **54**, 42-56.
- SHEPARD, R. N., & COOPER, L. A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.
- SHEPARD, R. N., & METZLER, J. (1971). Mental rotation of three-dimensional objects. *Science*, **171**, 701-703.
- SERLING, G. (1960). The information available in brief visual presentation. *Psychological Monographs*, **74**(11, Whole No. 98), 1-29.
- TARR, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, **2**, 55-82.
- TARR, M. J., & BÜLTHOFF, H. H. (1998). Image-based object recognition in man, monkey and machine. In M. J. Tarr & H. H. Bülthoff (Eds.), *Object recognition in man, monkey, and machine* (pp. 1-20). Cambridge, MA: MIT Press.
- TARR, M. J., & PINKER, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, **21**, 233-282.

- ULLMAN, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, **32**, 193-254.
- ULLMAN, S. (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 & Machine Intelligence*, **13**, 992-1006.
- WADE, N. J. (1973). Orientation and spatial frequency effects on linear afterimages: The retinal reference for selectivity—A supplementary report. *Perception & Psychophysics*, **14**, 384-386.
- WILSON, K. D., & FARAH, M. J. (2003). When does the visual system use viewpoint-invariant representations during recognition? *Cognitive Brain Research*, **16**, 399-415.
- WRAGA, M., CREEM, S. H., & PROFFITT, D. R. (1999). The influence of spatial reference frames on imagined object- and viewer rotations. *Acta Psychologica*, **102**, 247-264.
- WRAGA, M., CREEM, S. H., & PROFFITT, D. R. (2000). Updating displays after imagined object and viewer rotations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 151-168.

(Manuscript received January 20, 2004;
revision accepted for publication December 13, 2004.)