
Metric Invariance in Object Recognition: A Review and Further Evidence

ERIC E. COOPER, IRVING BIEDERMAN, and
JOHN E. HUMMEL *University of Minnesota*

Abstract Phenomenologically, human shape recognition appears to be invariant with changes of orientation in depth (up to parts occlusion), position in the visual field, and size. Recent versions of template theories (e.g., Ullman, 1989; Lowe, 1987) assume that these invariances are achieved through the application of transformations such as rotation, translation, and scaling of the image so that it can be matched metrically to a stored template. Presumably, such transformations would require time for their execution. We describe recent priming experiments in which the effects of a prior brief presentation of an image on its subsequent recognition are assessed. The results of these experiments indicate that the invariance is complete: The magnitude of visual priming (as distinct from name or basic level concept priming) is not affected by a change in position, size, orientation in depth, or the particular lines and vertices present in the image, as long as representations of the same components can be activated. An implemented seven layer neural network model (Hummel & Biederman, 1992) that captures these fundamental properties of human object recognition is described. Given a line drawing of an object, the model activates a viewpoint-invariant structural description of the object, specifying its parts and their interrelations. Visual priming is interpreted as a change in the connection weights for the activation of: a) cells, termed *geon feature assemblies* (GFAs), that conjoin the output of units that represent invariant, independent properties of a single geon and its relations (such as its type, aspect ratio, relations to other geons), or b) a change in the connection weights by which several GFAs activate a cell representing an object.

Résumé Sur le plan phénoménologique, il semble que l'être humain identifie les formes de façon invariante en dépit de changements d'orientation en profondeur (jusqu'à la dissimulation de parties), de position dans le champ visuel et de grandeur. Selon de récentes versions des théories des patrons (p. ex. Ullman, 1989; Lowe, 1987), il y a invariance de la forme lorsque des transformations sont effectuées, par exemple une rotation, une translation ou encore une mise à l'échelle de l'image de sorte que celle-ci peut être associée de façon métrique à un patron stocké en mémoire. De telles transformations nécessiteraient vraisemblablement

blement un délai d'exécution. L'étude porte sur de récentes expériences d'amorçage dont le but était de déterminer dans quelle mesure les sujets pouvaient reconnaître une image après l'avoir vue brièvement. Ces expériences ont démontré que l'invariance est complète: un changement de position, de grandeur, d'orientation en profondeur, ou encore les lignes et les vertex contenus dans une image, n'influent pas sur l'ampleur de l'amorçage visuel (distinct de l'amorçage engendré par le nom ou par le concept de base), pourvu que les représentations des mêmes composantes puissent être activées. L'article fait état d'un modèle de réseaux de neurones constitué de sept couches (Hummel et Biederman, document sous presse) et mettant en oeuvre les propriétés fondamentales qui entrent en jeu dans l'identification des objets par l'être humain. Ce modèle déclenche une description structurale d'un objet, représenté par un dessin au trait, description qui est invariante quel que soit le point de vue et qui illustre les parties de l'objet ainsi que leurs relations. L'amorçage visuel est interprété comme une modification des poids de connexion permettant d'activer: a) des cellules, appelées assembleurs de caractéristiques de géon, qui relient la sortie d'unités représentant les propriétés invariantes et indépendantes d'un seul géon et ses relations (p. ex. type, rapport de forme, relations avec d'autres géons) ou b) une modification des poids de connexion par lequel plusieurs assembleurs de caractéristiques de géon activent une cellule représentant un objet.

The nature of the memory representation used for visual object recognition is a central problem for students of cognition. Humans show a remarkable ability to rapidly recognize objects even though they may appear at an infinite number of sizes, orientations in depth, and positions in the visual field. Two classes of theoretical accounts have been proposed regarding how a common memory representation can be activated given such a wide variety of viewing conditions.

One class holds that the memory representation used for object recognition is *metrically specific*. That is, the representation is specific with respect to size, position, and orientation. Matching a new input (also at a specific size, position, and orientation) to a memory representation therefore requires that one or the other undergo a transformational process in which these metric properties are brought into congruence before the memory representation can be activated. Ullman's (1989) Feature Alignment Model and Lowe's (1987) SCERPO model, for example, use such an approach.

Alternatively, the memory representation may be *metrically invariant*. That is, the representation could be activated by the object no matter its size, position, or orientation in depth. Object recognition theories of this type, such as Marr & Nishihara (1978) and Biederman's (1987) Recognition by Components (RBC), tend to be structural description theories (Pinker, 1984) in which the object's "parts" and their spatial relations are represented categorically.

Insofar as shape must be the output of a system that is sensitive to *metric* variations in orientation and position, as would be computed, for example, by simple cells in V1, these metrically invariant theories typically assume metric processing at early stages. However, once the representation of shape is derived, it is assumed to be independent of the particular metric sensitive units that produced the shape descriptors.

A distinction must be drawn between representations for recognition and representations subserving motor interactions with objects. The former may correspond to the perceptual representations of the ventral visual system and the latter, the dorsal visual system (Mishkin & Appenzeller, 1987; Biederman & Cooper, 1992). We assume that motor representations are metrically specified, as competent interactions require sensitivity to where an object is, its size, and its orientation. We devote some space in this article to distinguishing empirical effects that could arise from one or the other system.

A Critical Review of the Evidence for Metric Specificity

Previous empirical investigations as to whether or not a metric transformation process occurs during object recognition would seem, at first glance, to support metric specificity. A number of studies have shown time and error costs for matching shapes that differ in their depth orientation (e.g. Bulthoff, Edelman, & Sklar, 1991; Shepard & Metzler, 1972; Tarr, 1989), with the magnitude of these costs increasing directly with the degree of disparity in orientation.

Other results that have been cited as supportive of metric specificity derive from simultaneous or immediately sequential shape matching experiments, in which subjects are required to judge if a pair of shapes are the same or different, using stimuli that differ in size. Such experiments have generally found time costs for performing the task varying directly with the size disparity (e.g. Besner & Coltheart, 1975; Bundesen & Larsen, 1975; Bundesen, Larsen, & Farrell, 1981; Howard & Kerst, 1978; Jolicoeur & Besner, 1987; Larsen, 1985; Larsen & Bundesen, 1978). The general explanation for these results is that the representations used for object recognition are stored at a particular size with a time consuming size scaling procedure necessary for object recognition to occur.

The relevance of these results for *basic-level object classification* (rather than shape *discrimination*) may be questioned, however. Just because an experiment employs visual stimuli that vary in shape does not automatically mean the results of that experiment should be generalizable to the perceptual classification of shape into basic level categories. (Actually, in the present context, *entry level*, rather than basic-level, would be the more appropriate term in that penguins and ostriches would qualify for their own class, separate from that for bird, as argued by Jolicoeur, Gluck, & Kosslyn, 1984, and Biederman, 1987).

The criteria by which experimental data are judged to be relevant to a given real-world activity depend, of course, on one's theory of how that activity is performed. From the standpoint of RBC, the experiments purporting to support metric specificity lack relevance to object recognition in that: a) they employed tasks that might not have reflected the critical stages of object classification, or b) they employed sets of stimuli in which the individual members were not distinguishable by the kinds of stimulus attributes that characterize naturally-occurring basic level classes (which, in turn, may be determined by a specific set of perceptual processes).

With respect to the first factor, some of the tasks possibly allowed processing to be completed before stages presumed critical to object classification could be engaged. An example of such a task would be one in which subjects performed same/different shape matching of a pair of simultaneous or sequentially presented images (e.g., Bundesen & Larsen, 1975; Howard & Kerst, 1978; Larsen, 1985). Other tasks may have been susceptible to the influence of processes other than those involved in object classification. For example, the feelings of familiarity that determine episodic recognition memory judgements ("Did you see that shape in the prior block of trials?"), as in the experiments by Jolicoeur (1987) and Biederman and Cooper (1992, Experiment II), may be influenced by a system also involved in the visual memory supporting motor interactions that might represent the location of objects in the environment and their sizes (Biederman & Cooper, 1992).

Perhaps the effects of orientation or size differences on simultaneous or immediately sequential same/different judgments represent influences on a visual stage that occurs prior to object recognition. This stage might facilitate discriminations between objects currently present in the visual field, but it may not be generally available when an image must be compared to a memory representation as is required for entry level object recognition. Support for this notion comes from Ellis and Allport (1986) who found an effect of rotation in a same/different shape matching task using sequentially presented real objects at an ISI of 100 msec. However, when the ISI was increased to 2000 msec, the rotation effect disappeared.

The second factor restricting the generalizability of previous studies to basic level object recognition concerned the stimulus attributes that distinguish a set of stimuli. Not all distinctions among shapes are the kind across which cultures develop entry level categories. In particular, a number of experiments in which an effect of rotation angle was obtained employed sets of stimuli that could not be distinguished by the kinds of viewpoint invariant contrasts posited by RBC. The stimuli, in some experiments, were not readily decomposable into invariant parts and relations as, for example, the smoothly curved wire stimuli in the Rock and DiVita (1987) experiment and their clay molds and crumpled newspaper demonstrations. With the wire objects, loops

(that, subjectively, are the prominent parts for these shapes) are accidents of viewpoint in that they are apparent in one view but disappear in another, despite the presence of their component contours in the image. In other experiments, the stimuli do decompose into invariant parts and relations, but the members of the set of stimuli all have the same *pairwise* part-relations descriptions. For example, Tarr (1989) used stimuli that were arrangements of five or six bricks made up of varying numbers of cubes. These stimuli were readily decomposable into parts (the bricks) and relations (right angle end-to-end or end-to-middle joins). (Moreover, they all had a three-cube brick as a foot and a long brick attached at one end to the middle of the foot.) But members of the set of seven stimuli could not be distinguished on the basis of individual part-relation combinations (e.g., vertical cylinder ON-TOP-OF; horizontal brick BELOW) of the type represented in the neural net implementation of RBC (Hummel and Biederman, 1992 described in a subsequent section). The distinguishing information for the Tarr (1989) stimuli required what can be termed *third order descriptors*: Not just a geon and a relation must be specified, but an additional property as well. In this case it is the metric specification of the lengths of pairs of connected bricks, so a (verbal) distinguishing description for stimulus two in Tarr (1989) would be: one of the two longest bricks connected end-to-middle to a three-cubed brick *and* the other long brick connected end-to-end to a two-cubed brick. It is unlikely that any culture would form entry level classes from such third order descriptions. Indeed, the Hummel and Biederman (1992) neural net implementation of RBC would have great difficulty in distinguishing among members of such classes. When entry level object classes are composed of the same geons, as with a cup and a bucket, for example, they will differ in the types of relations or parts. Indeed, in contrast to the Tarr (1989) and Rock and DiVita (1987) results showing marked effects of orientation in depth in shape matching tasks, Gerhardstein & Biederman (1991) reported extremely small effects of differences in rotation angle when using a set of "nonsense" objects whose members shared the same geons and relations but that could be distinguished according to the relations bound to a given geon.

The purpose of this paper is to review a number of experiments conducted in our laboratory that provide evidence for the existence of metrically invariant shape representations for purposes of recognition and to present new experimental evidence relevant to this issue. Further, theoretical and empirical work suggesting the form of the representation that allows for this metric invariance will be discussed.

Evidence for Metric Invariance from Priming Experiments

The general research paradigm we have used to study metric invariance is the *priming paradigm*. The speed and accuracy of naming object pictures has been shown to improve markedly with a second presentation of the pictures

(e.g., Bartram, 1974; Biederman & Cooper, 1991a,c; 1992). In a study of size invariance, Biederman & Cooper (1992) presented line drawings of common objects displayed at one of two sizes, small (3.5 degrees of visual angle) or large (6.2 degrees). After presentation of a first block of trials, subjects viewed a second block of objects all of which had the same names as those presented in the first block. In all the priming experiments reviewed here, second block RTs and error rates were markedly lower than they were on the first block, indicating that priming had occurred. (General learning-to-learn effects that could have mimicked priming were ruled out in a control experiment, Biederman & Cooper, 1991a).

Half the images in the second block were of a different shaped exemplar of the object class with the same name. For example, if a grand piano were presented in the first block, an upright piano presented in the second block would be a *different exemplar* (see Fig. 1). Objects in the second block were either the same or the other size relative to what they were in the first block.

The different exemplar condition served as a control to ensure that a component of the priming was, indeed, visual. A visual component of the priming would favour performance on the same-exemplar trials compared to the different-exemplar trials. In fact, any difference in priming observed between same and different exemplars likely *underestimates* the amount of visual priming that was occurring. Visual similarity between two exemplars of the same class was, in general, greater than that across classes. For example, in the images that were used, both birds had a beak and wings; both cars had wheels, windows, and doors. Given that a difference was observed between same and different exemplar trials, the critical comparison for determining size invariance is between the same and different size conditions for same exemplars. If a time consuming transformational process was required to match a new input to an old representation, then less priming would have been expected for the different size condition corresponding to the time required to perform the transformational process. If, on the other hand, the objects are represented in a manner that is invariant with respect to size, then equivalent priming for same and different size conditions would be predicted.

Biederman & Cooper (1992, Experiment I) found considerable priming between first and second blocks with different exemplars in the second block showing significantly less priming than same exemplars (thus providing evidence that a portion of the priming was visual). Most importantly, equivalent priming was observed for the same and different size conditions indicating that the representation used for object recognition is size invariant (Biederman & Cooper, 1992, Experiments I and III).

To assess whether the sizes used in this experiment were sufficiently disparate to show an effect, Biederman & Cooper (1992, Experiment II) used the presentation conditions of the priming experiment to replicate the Jolicoeur

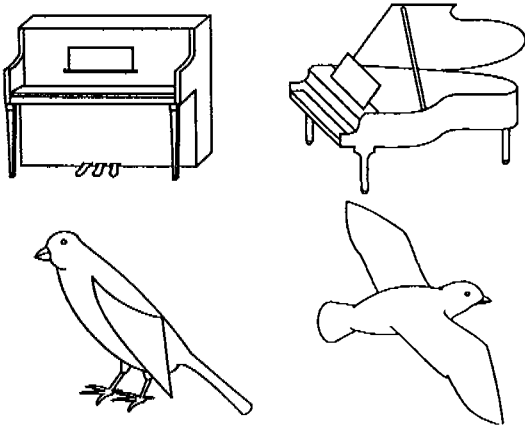


Fig. 1 Illustration of different exemplars for the classes piano and bird. Modified from Biederman, I., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 121-133. Reprinted by permission of the authors and the American Psychological Association.

(1987) results, in which costs were observed for size changes from study to test in an episodic memory task. In the replication, subjects named objects in a first block of trials (exactly as in the priming experiments) but in the second block, they were required to say whether the object viewed was the “same” or “different” in shape from those seen in the first block ignoring any changes in size. The “different” trials consisted of different exemplars of objects seen in the first block. (This was not the case in the Jolicoeur experiment in which the “different” trials were for images from categories [names] that did not appear on the first block.) The results replicated those of Jolicoeur (1987) with considerable costs in both response times and errors for “same” responses when size changed from first to second block.

Why were the effects of size changes not observed on naming but were present when judging same/different shape? Biederman & Cooper’s (1992) interpretation was that the episodic judgements reflected influences from two different memory representations. One representation mediate object recognition and was assumed to be size invariant. The other representation, perhaps mediating motor interaction, was presumed to be size specific (or at least specified a binding of the size information to the shape of the object). The naming task, which showed equivalent priming, required only access to the shape representation. To perform the episodic memory task, however, subjects relied on a feeling of overall familiarity (Atkinson & Juola, 1974) that could have been affected by both size-invariant and size-specific representations. Subjects were able to respond “same” quickly and accurately in the episodic memory task when the size remained the same from first to second block because such a stimulus would have provided an identical match to both representations, thereby giving a strong sense of familiarity. If the size

changed, in contrast, the feeling of familiarity would not have been as high, leading to slower, less accurate responses. Consistent with this interpretation that naming can be performed from the activation of only a single representation rather than through a build up of episodic information from several systems, was that naming, despite its considerably higher response uncertainty and response production demands, showed itself to be a much easier task than episodic judging, with markedly lower second block RTs and error rates than that for making episodic memory judgements.

Given that the task required of a subject in the episodic memory paradigm would be best performed by relying only on the size-invariant representation (because size was irrelevant), it may seem puzzling that the size-specific representation would be accessed as well when performing the task. Perhaps the representations accessed in order to create a "feeling of familiarity" were not under the subject's volitional control, but rather were automatically pooled. This sense of familiarity served as the subject's primary means of determining whether a specific instance of an object had been seen before. In contrast, the sense of familiarity for a particular object was irrelevant when a subject was required to perform basic level classification of objects, and the name of the object can be accessed from the size-invariant representation without the influence of a size-specific representation. Recall that in Biederman and Cooper's (1992) version of the episodic memory task (in which costs were observed for making old-new judgements after object size changed from first to second presentation), subjects simply had to *name* objects on a first block of trials prior to making an old-new judgement in a second block. This result suggests that even in a naming task both size-invariant and size-specific representations are activated. Thus, perhaps, the representations that are *activated* do not differ depending on whether a naming or episodic judgement task is performed, but rather, how those representations are *accessed* changes.

Using the identical priming paradigm, the metric invariance of the representation used for object recognition have been confirmed for other dimensions. Biederman & Cooper (1991b) found equivalent priming for pictures that were mirror-image reflected from first to second block, and for pictures that had been translated in the plane both horizontally and vertically. Time costs for translations from first to second block in an episodic memory experiment analogous to the one described earlier for size have also been observed (Cooper & Biederman, 1992a). Further, Gerhardstein and Biederman (1991) had found equivalent priming for images rotated in depth from first to second presentations provided that the same convex components are in view in both images.

Note the methodological advantages of experiments using the priming paradigm over simultaneous (or immediately sequential) same/different shape matching and episodic memory experiments for assessing the memory

representation used in object recognition. In contrast to simultaneous or immediately sequential matching tasks (e.g., Bundesen & Larsen, 1975; Howard & Kerst, 1978; Larsen, 1985), the length of time between the first and second presentations of the stimuli using a priming paradigm (generally between five and ten minutes in our experiments) precludes the use of a short term visual buffer for stimulus matching. Priming does not require the subject to explicitly remember a previous encounter with a stimulus, thus, perhaps, allowing a purer measure of shape representation activation than episodic memory paradigms (e.g., Jolicoeur, 1987) which may possibly be tainted by the influence of other, independent representations.

The Nature of the Representation

How can object shape be represented in such a way that the representation can be activated by an object that may vary along a number of metric dimensions? A number of theorists (e.g. Biederman, 1987; Brooks, 1981; Marr & Nishihara, 1978; Palmer, 1975, 1977; Tversky & Hemenway, 1984) have argued for a representation based on an object's parts. For example, RBC proposes that object recognition proceeds by extraction of *image features* from an object, separation of the object into its *convex or singly concave components*, identification of each component as a particular member of a small set of geometric primitives (geons), and finally, activation of an *object model* based on the identity of the components and their relations to one another. Because the primitive components are robust to noise, orientation change, and size change, such a representational scheme allows recognition of objects under wide metric variations.

Biederman and Cooper's (1991a) attempt to find the representational level at which priming occurs provides strong evidence that the representation used for object recognition does consist of a specification of the object's parts and their relations. They created stimuli by deleting every other image feature (edge and vertex) from line drawings of objects in order to create two *complementary images* of each object. That is, the two images for each picture shared no common contour, and when placed together, formed the complete object (see Fig. 2). The complementary images were created in such a way that each convex component of the object could be recovered from each of the images. Thus, although complementary images shared no edges and vertices, they shared the same components. Because the amount of contour deleted from each image was substantial and included vertices, it was unlikely that a local process of filling-in to complete the contour occurred with these images (see Biederman & Cooper, 1991a, for a more complete discussion).

These feature-deleted stimuli were used to determine whether perceptual priming in object naming takes place at the level at which image features are processed. Using the priming paradigm discussed earlier, subjects named the

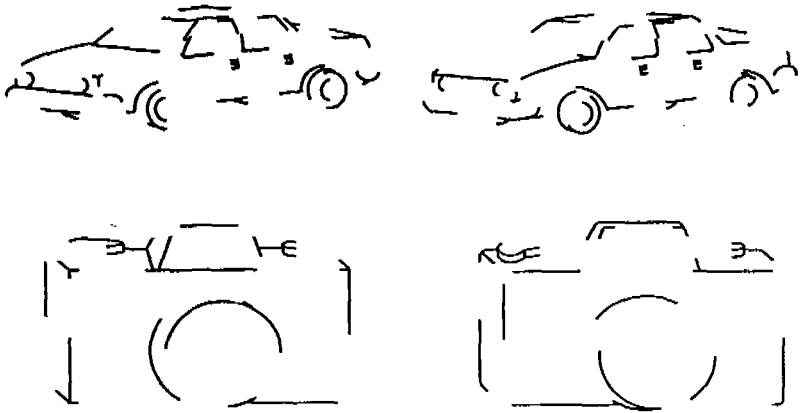


Fig. 2 Illustration of a complementary feature pair for two object classes, a car and a camera, for the Biederman and Cooper (1991a) complementary feature priming experiment. Alternate vertices and lines have been removed from each part. Long contours were divided in half so they appeared in each of the members of a complementary pair. Images from a complementary pair, when superimposed, would make an intact image with no overlap in contour.

feature-deleted stimuli on a first block of trials. On each trial in the second block, subjects could see the identical picture they had seen in the first block (with the same contours deleted), the complement of a first block picture (with the remaining image features) or a different exemplar of a picture they had seen in the first block (also with contour deleted). The results showed substantial priming for the second block trials, with different exemplar pictures significantly slower than both identical and complementary pictures - thus the priming did include a visual component. The key result of this experiment was that identical and complementary pictures showed *exactly the same amount of priming*; thus the particular image features present in an object do not mediate perceptual priming.

Adopting Biederman's (1987) theoretical perspective, this experiment eliminated image feature processing as the locus of priming, but still allowed that visual priming could take place at the level where the object's components are activated or at the level where the object model is activated. To decide between these alternatives Biederman and Cooper (1991a) repeated their experiment using complementary images that had been created by *deleting entire convex components* from line drawings (see Fig. 3). Note that complementary images created in this way shared the same object model, but have neither components nor image features in common.

As with the feature-deleted stimuli, significant priming was evident from the first to the second block with different exemplar trials showing markedly less priming than identical trials. In striking contrast to the feature-deleted stimuli, however, the complementary condition with component deleted

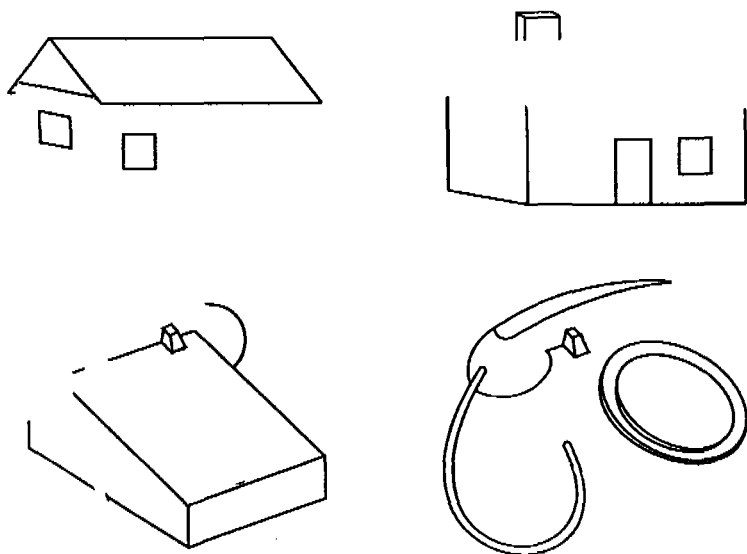


Fig. 3 Illustration of a complementary pair for two object classes, the house and telephone, for the complementary parts priming experiment. Images form a complementary pair that, when superimposed, would make an intact image with no overlap in contour.

stimuli showed priming *identical to that of the different exemplar condition* (significantly less than that for the identical trials). Thus, when different components are present in two objects that share the same object model, the perceptual portion of priming is completely eliminated.

Probing the Limits of Invariance: Priming Translated, Reflected, Countour-Deleted Stimuli

The conclusion from the experiments reviewed here was that shape priming occurs to the degree to which neural units, termed *geon feature assemblies* (GFAs) by Hummel and Biederman (in press), are activated. GFAs represent a particular geon (including its aspect ratio and orientation) and its relations to other geons. This point will be extensively discussed in the next section. The main point for the present consideration is that the units are assumed to be invariant to translation, scale, and orientation in depth, as well as to the particular local contours that activated the geons.

The previous section described the results from several experiments documenting these invariances. There is some evidence that simultaneous changes in two of the variables also have no effect on the magnitude of priming. In the Biederman and Cooper (1991a) experiment with feature-deleted stimuli, half the stimuli on the second block (both identical and complements) were presented in their original orientation and half were

presented in mirror-image reversed orientation. There was no reduction in priming as a result of the change in orientation. Nor was the magnitude of priming reduced by a change in orientation for intact, translated stimuli in the Biederman and Cooper (1991c) experiment. In that experiment, stimuli were presented either to the left or right of fixation. On the second block, the objects could be presented in either the same or different visual hemifield, at the same or different left-right orientation.

We report here an additional experiment designed to provide an assessment of the limits to the invariances evidenced in the prior experiments. In this experiment, one member of a feature-deleted complementary pair of images (as in Biederman & Cooper, 1991a) was presented either on the left or right side of the screen in the first block of trials. In the second block, subjects saw an image identical or complementary to what they had seen on the first block, at the same or different position on the screen. (Given that every experiment we have conducted has shown substantially more priming for same exemplar than different exemplar trials, this manipulation was not included.) Subjects were divided into two groups. One group, the *constant orientation* group, saw every object in the same orientation on the second block as it had appeared on the first block, independent of whether it changed position. The other, *mirror-image* orientation group would see objects in the original (first block) orientation when their position did not change from first to second block but in mirror image orientation when the position did change.

From the perspective of the prior analyses, there were two likely outcomes. If both constant and mirror reversed conditions activate the same GFAs, then no differences would be expected from a reversed stimulus. But the displacement of contour deleted images away from central fixation would increase the chance that some of their parts might not be recovered to activate the GFAs. In particular, the more eccentric the part, the less likely it would be to be detected. This analysis leads to a somewhat counterintuitive prediction as to whether an orientation change would affect the magnitude of priming: If a difference is observed in the magnitude of priming of same and mirror reversed translated images, it should be the mirror reversed stimuli that show the greater degree of priming. For example, when presented on one side of the screen, the dog would have its head and front legs close to the fixation point, while on the other side of the screen in the consistent orientation group, the tail and hind legs would be close to the fixation point. On the other hand, exactly the same components will be close to the fixation point at both positions for the mirror-image orientation group.

METHOD

Subjects

The subjects were 64 native English speakers with normal or corrected-to-normal vision. They participated for payment (\$5 per session) or research

experience points for the Introductory Psychology course at the University of Minnesota.

Stimuli

Thirty-two line drawings of common objects and animals were used as stimuli. For each picture, two complementary versions were created by deleting every other edge and vertex from the complete drawing. Thus, the two versions each contained roughly half the contour of the complete image with no overlapping contour between them. Further, the two complements were created in such a way as to make each convex component of the drawings equally recognizable in each image. Complete rules for creating the complementary images can be found in Biederman and Cooper (1991a).

The drawings were created in Cricket Draw (using a line width of two pixels) and shown on a high resolution (1024×768) monitor (Mitsubishi model HL6605) controlled by a Macintosh II. Each picture was scaled so that its maximum extent would just fit inside a circle with a diameter subtending 4 degrees of visual angle. The images were centered 2.4 degrees to the left or right of fixation, and so the closest possible point to fixation of any picture was .4 degrees.

Procedure

To initiate each trial, subjects pressed a mouse button. A central fixation dot would then be presented for 500 msec, followed by the 150 msec presentation of the object picture (a duration too brief for the subject to make a second eye fixation). After the picture, a mask (a dense arrangement of random lines) was presented for 500 msec. Subjects were instructed to maintain fixation on the center of the screen throughout the trial. The subjects' task was to name each object picture with its appropriate basic level term (e.g., "elephant") as quickly as possible. Naming reaction times were recorded using a Scientific Prototype voice key and errors were recorded by the experimenter. After each trial, subjects were given response time and accuracy feedback and, after a delay of one second, were given a message telling them they could press the mouse button to initiate the next trial.

A response was recorded as an error if it was not the basic level name of the object presented. False starts and responses that occurred more than three seconds after the object was displayed were also considered errors. Responses synonymous with the feedback term (such as "automobile" for "car") were not counted as errors.

Prior to the experimental trials, subjects were given 12 practice trials using pictures not presented in the experiment. The stimuli were presented in two blocks of 36 trials (two pictures at the beginning and end of each block were used as "buffer" trials on which no data were collected). All 32 pictures were presented in each block with the side of the screen on which each object was

presented varying randomly from trial to trial and the sequence of images randomly selected for each block. Approximately 5 minutes intervened between presentations of an object in the first block and presentation of its partner in the second.

Design

Subjects were divided into two groups of 32, a *constant orientation group* in which the orientation of the objects remained the same no matter which side of the screen the object appeared, and a *mirror-image orientation group* in which the orientation of an object when it appeared on the left side of the screen would be the mirror image of when it appeared on the right. For example, the constant orientation group always saw the fish pointed with its head to the right, while the mirror-image orientation group saw the fish pointed to the right when it was on the left side of the screen and pointing left when it was on the right side. Thus, object orientation was not balanced for the constant orientation group and was confounded with position for the mirror-image orientation group. Whether the object was facing left or right was selected randomly for each object. Biederman & Cooper (1991b) found no effect of their objects' left-right orientation on naming latencies.

The objects in the second block of trials could appear in either the identical position in which they appeared on the first block or the opposite position (on the other side of the screen). Further, in the second block the object could have exactly the same image features presented as in the first block or the complementary features. Thus the variables of interest in the experiment were in a 2 (constant orientation vs. mirror image orientation) \times 2 (same position in the second block vs. different position) \times 2 (identical image vs. complementary image) design.

The stimuli were balanced in such a way that each subject saw an equal number of objects in each Position (same-different) \times Image (identical-complement) cell. Across subjects each object appeared equally often in each cell. Half the subjects saw the stimuli presented in forward order and half in reverse, thus balancing the mean serial position of every object. Two subjects in each of the orientation groups would see exactly the same objects in exactly the same conditions (one in forward order and one in reverse).

ANOVAS for both RTs and errors were computed. In these analyses, Orientation Group (constant vs. mirror image), Position (same vs. different), and Image (Identical vs. Complement) were included as fixed factors with subjects as a random factor.

RESULTS

The mean correct RTs and error rates for all 64 subjects are shown in Figures 4 and 5, respectively, pooled across orientation group. Significant priming from the first to the second block was evident. Mean RTs and error rates were

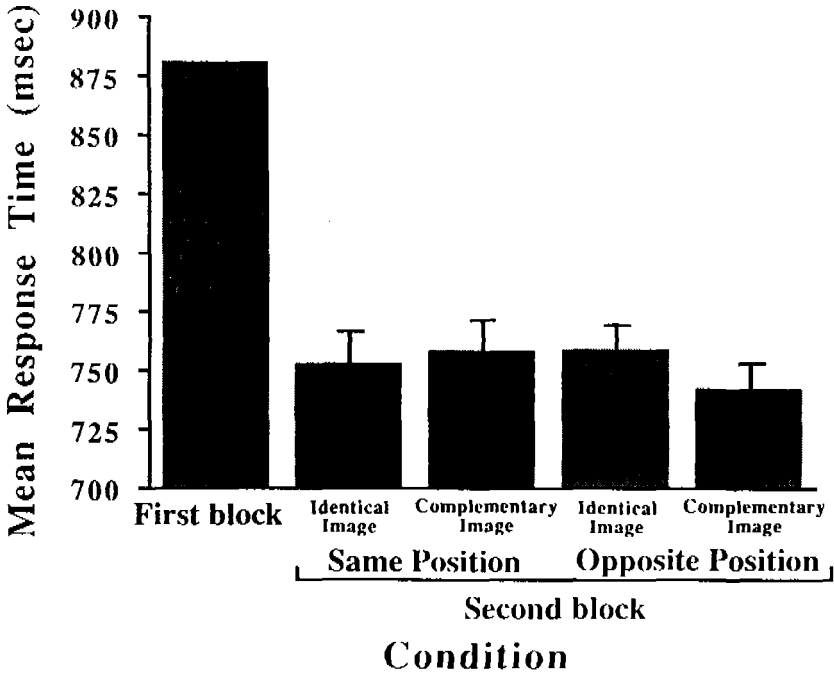


Fig. 4 Mean correct naming reaction times (RTs) for first and second blocks in an experiment that varied Position, Image Type, and Orientation. Second block data are shown for the effects of changes in Image (Identical or Complement) and Position (Same or Opposite). The data are collapsed over Orientation Groups. Second block data are for those trials where the object was correctly named on the first block. (Inclusion of those trials where the first block was in error did not alter the pattern of the results, although it did increase variability.) Error bars on the second block are the standard errors of the difference scores between each subject's mean score on block 2 and his or her mean score for that condition.

sharply lower on the second block than they were in the first for both groups, both $ps < .001$.

None of the main effects for RTs were close to significance (all $F_s < 1.00$). There was, however, a significant effect of position for the error rates: A change in position produced a higher error rate, $F(1, 62) = 8.60, p < .01$. Although the interaction term for Groups \times Position, fell short of significance, $F(1, 62) = 2.40, p < .13$, almost all this effect came from the Constant Orientation group. A change in position for this group produced an increase in error rates from 2.2% to 8.0%. The corresponding increase for the Mirror-Reversed Group was smaller, from 5.7% to 7.5%. None of the other main effects or interactions were significant at $\alpha = .15$.

DISCUSSION

For the most part, the results of this experiment document a striking degree

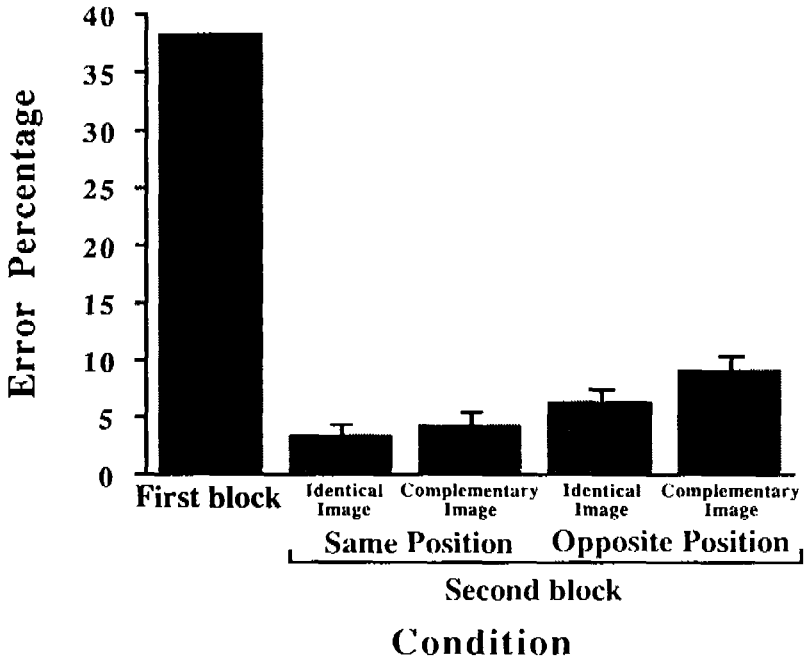


Fig. 5 Mean error rates for first and second blocks in an experiment that varied Position, Image Type, and Orientation. Second block data are shown for the effects of changes in Image (Identical or Complement) and Position (Same or Opposite). The data are collapsed over Orientation Groups. Second block data are for those trials where the object was correctly named on the first block. Error bars on the second block are the standard errors of the difference scores between each subject's mean score on that block and his or her mean score for that condition.

of invariance: A translated, reflected, complementary image suffered no loss in its identifiability, as assessed by priming, compared to an image that retained its original contours, in the same position, and in the same orientation. The suggestion of a departure from invariance for the group with the constant oriented pictures was consistent with the possibility that that condition would have been more susceptible to different geons being detectable when translated. Further, these results replicate those of Biederman & Cooper (1991a) in which reflectional invariance with similarly degraded stimuli was found.

What does it mean to say that priming is visual?

The purpose of this section is to consider these results in the context of an overall theory of object recognition, the neural net implementation of RBC (Hummel & Biederman, 1992). The model is a seven layer network that takes as input a line drawing representing the orientation and depth discontinuities

of an object and, as output, activates a unit representing a particular object. Figure 6 shows the model's overall architecture. Within the context of the model, the question posed at the heading of this section can be reformulated to be: At which layer(s) does visual priming occur? What is the mechanism for priming? We pose these questions in the context of this particular model, not because we believe that it has been shown to be absolutely correct, but because it provides the most complete account currently available of real-time human object recognition.

Overview of the model

The first layer can be regarded as a highly simplified V1. It consists of 484 identical columns (analogous to V1 hypercolumns), each with 48 cells that are selectively tuned to the orientation of an edge and whether the edge: a) is curved or straight, and b) extends through or terminates within the limited receptive field of the column. The receptive fields of adjacent columns overlap, so that edges are coarsely coded by several cells. The degree of activation of a cell provides a measure of the extent to which a cell's feature is present in the receptive field. Grouping, or binding, is done through *fast enabling links* (FELs) that cause cells that are simultaneously active to fire together if their receptive fields are cocircular (or collinear), cotermine, or are closely parallel. The FELs produce synchronous firing of all the cells that are activated by a given geon while allowing cells activated by different geons to fire out of phase with each other.

In the model's second layer are three sets of cells that represent vertices, axes, and blobs, again at particular locations in the visual field. The vertex cells receive their input from the output of the termination cells in layer 1. The local units of the second layer activate units in the third layer that represent viewpoint-invariant properties of geons (cross section curvature, axis curvature, and whether the sides are parallel), and coarsely coded representations of its size, position in the visual field, aspect ratio, and orientation. These latter metric properties are used by layers 4 and 5 to derive the relations among the geons in an image: relative size (e.g., LARGER THAN), relative position (e.g., ON TOP OF), and relative orientation (e.g., PERPENDICULAR TO).

The geon attributes and relations are termed "invariant" in that each unit will respond to a geon with its target property (e.g., curved axis) regardless of the geon's other properties.

The sixth layer consists of units that self-organize to patterns of activation over the geon units of layers three and the relations units of layer five. These are the GFAs described earlier. For example, a given L6 unit might respond to a cylinder (actually cells indicating a curved cross section, straight axis, and parallel sides), vertically oriented, and above, smaller than and perpendicular to something else. Each of an object's geons will activate a different cell in L6.

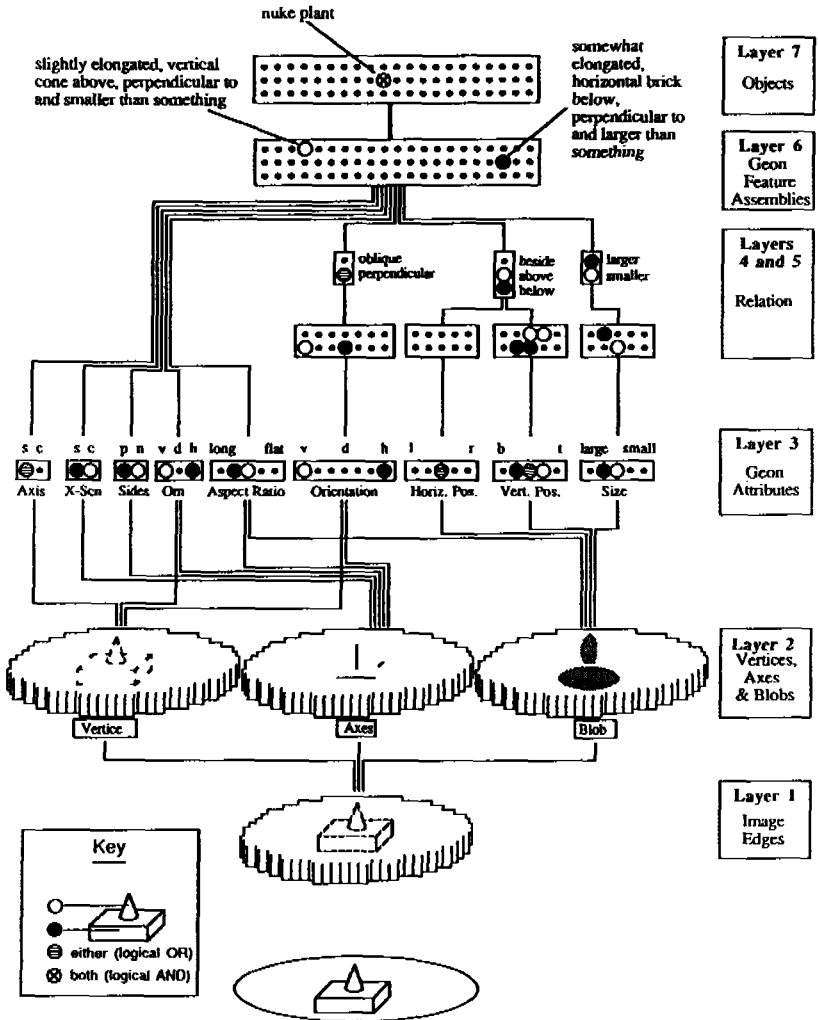


Fig. 6 The architecture of the neural net implementation of RBC. Figure modified from Hummel, J. E., and Biederman, I. *Dynamic binding in a neural net model for shape recognition*, *Psychological Review*, 1992. Reprinted by permission of the authors and the American Psychological Association.

The "object" cells in layer seven self organize to respond to conjunctions of L6 cells. A unit in this layer might thus represent an object consisting of a vertical cylinder centered above and smaller than a horizontal brick.

Locus and Mechanism of Priming

We now return to the original problem addressed at the outset of this section as to the locus and mechanism of priming.

Priming did not occur at layers that represent local image properties. Units in the model's first two layers respond to *local* inputs. Changing the size, position, or orientation of the image would activate a completely different set of units in the first two layers. That the magnitude of priming was not reduced by such changes is strong evidence against the locus of priming at units with spatially restricted receptive fields. Moreover, the equivalence in priming between complementary and identical feature-deleted images (Biederman & Cooper, 1991a), in which alternative vertices and edges were deleted from each geon, provides strong evidence against such an account. Members of a complementary pair of feature-deleted images would activate not a single common unit in the first layer, nor any common vertex units in the second layer.

Priming did not occur at units representing single, invariant attributes (layers 3, 4, and 5). In these experiments, subjects viewed multiple objects over several minutes, so the same cells in layers 3, 4, and 5 would have been activated many times over. For example, a number of different objects had parts with curved cross sections, or parts that occupied common positions in the visual field, or were elongated, or had the relation ON TOP OF, etc. Could priming be a function of persistent, heightened activation of these units (or an increase in the connection weights by which activation from earlier layers was passed to these units)? Most likely not. Increased activation of these units would be associated with a large number of objects (not just on the experimental screen but in the room as well). Any gain from the increased activation would have to be balanced against the interference from activating a large number of different objects.

It seems unlikely that it would be advantageous for a perceptual system to maintain a bias in favour of the kinds of information represented in the first five layers of the model. If a straight line or an L vertex or a brick or something ON TOP OF something else is detected in one scene, there would be no reasonable advantage in the system's adjusting its parameters to detect more of the same.

The preceding two paragraphs presented a plausibility argument against the locus of priming at the earliest stages of perceptual processing. Cooper and Biederman (1992b) performed an experiment designed to assess whether priming could occur from a single part presented immediately before an object. Each trial consisted of a rapid sequence of four stimulus events: a) a presentation of an object part (or control line) for 100 msec, b) a 50 msec mask (random appearing lines), c) a 100 msec presentation of a picture of an object, and d) a 500 msec mask. The subjects were to name the picture as quickly as possible. On half the trials when a part was presented, it was contained in the object. On the other half of the trials, the part was altered by making curved lines straight or straight lines curved (thus changing the geon categorization of the part). The parts, control line, and objects were all

centered at the fixation point, so the position of the parts did not correspond to their position in the object. Original parts, altered parts, and control lines were presented with equal frequencies (each for one third of the trials). For both naming RTs and error rates, no difference among the three priming conditions was obtained. It was not that the experimental paradigm did not allow priming: Naming RTs and error rates were markedly lower for a group of subjects who were shown (and named) the complete objects prior to the experiment. But these subjects, as well, did not evidence any difference in the priming conditions.

Priming is a modification of the connection weights to the Geon Feature Assembly layer (layer 6) and/or to the Object Layer (layer 7). If the mechanism of priming is a modification of connection weights between layers, the most plausible loci for priming would be in the connection weights to the units in the last two layers. These are the only layers in which learning occurs (excepting the kind of developmental or evolutionary learning that might have been required to form the first five layers).

One locus of priming would be in the modification of the connection weights between the pattern of activation of the invariant cells in L3 and L5 and the GFA layer (L6). At the GFA layer, activation specifying a given geon's attributes *and* its relations to other geons is, for the first time, used to organize a unit to represent the *combined* (or *anded*) activity from different invariant cells. The equivalence of priming for complementary and identical features (edges and vertices) suggests that not only was there no contribution from the local feature cells, but that the same geons and relations were activated for complementary pairs and, consequently, the same GFA units were activated.

Another way in which priming effects could be accommodated by the model would be in a lowering of the thresholds for the units in L6 (or in sustained activation of the units). Although a lowered threshold would be a possibility with a prime that immediately preceded a stimulus, the interval between *prime* and *stimulus* in our experiments was several minutes. With that length of time and so many intervening trials, a heightened state of excitation or lowered threshold would seem implausible. Moreover, biased excitation states or thresholds could produce errors given the high response uncertainty that was characteristic of these experiments. Experimentally, the data are inconsistent with lowered thresholds of the object cells in L7 as the locus of priming in that priming between complements composed of different convex components of an object was not greater than the priming between one member of that pair and a part complement of a different object with the same name (Biederman & Cooper, 1991a). If the threshold for firing a unit representing a grand piano would be lowered by seeing half of its parts, then the remainder of the parts of that grand piano should have produced more priming than half the parts of an upright piano.

This absence of priming of object units also suggests that there was no top-down facilitation and leads to the following strong conclusion from these experiments: *A distinctive GFA cell must be activated, bottom-up, for perceptual priming to occur.* Whether the priming occurred as modification of the connection weights (from the independent, invariant attributes and relations cells in layers three and five) to this cell or in the weights between the GFA cells and the object cells (or both) is a problem for future research. If the priming did occur in the connection weights to the GFAs, then a high "vigilance" value for activation of the GFAs must be assumed in which the relation cells (in layer five) must be activated. The reason for this is that the single geon primes specified the attributes for a single geon (such as aspect ratio and orientation) but did not specify the relations with other geons.

Physiological Support for Shape and Metric Attribute Independence

The notion of separate representations for shape for recognition and metric attributes for motor interaction receives some support from the results of cortical lesions studies. Lesions of the posterior parietal area have been shown to cause severe deficits in visual landmark tasks in which a monkey is required to make a response on the basis of a spatial cue (Pohl, 1973; Ungerleider & Brody, 1977; Brody & Pribram, 1978; Ungerleider & Mishkin, 1982) while tasks that require the discrimination of shape remain relatively unimpaired (Ungerleider & Mishkin, 1982). The converse effects are observed for inferior temporal (IT) lesions (Ungerleider & Mishkin, 1982; Mishkin & Appenzeller, 1987).

Patients with lesions to the superior parietal region have no trouble on recognition tasks or aligning a bar to a distant bar or judging the width of an object. However, when asked to interact motorically with those stimuli, by mailing a letter into a tilted slot or picking an object up, the orientation and size information that was available for perceptual judgements appears to be largely absent (Perenin & Vighetto, 1988). Quite the opposite pattern was characteristic of a patient with damage to the ventral visual system (Goodale, Milner, Jakobson, & Carey, 1991).

The results from the lesions studies are suggestive of different neural loci for metric and shape information storage, but this conclusion must be drawn with caution. Studies involving posterior parietal lesions have focused almost exclusively on deficits involving the spatial position of stimuli; whether these deficits extend to other metric attributes is still unknown. Further, the deficits resulting from posterior parietal damage have invariably involved the use of visual information for motor control. The possibility exists that, rather than being the site of memory for metric attributes, the posterior parietal area is only involved in using metric attributes for motor control and that memory for both shape and metric attributes *for recognition* are independent-

ly coded by the ventral cortical pathway ending in IT. Given that changes in metric attributes can have large effects on explicit episodic memory tasks (such as those of Jolicoeur, 1987), the size-specific representations cannot simply be part of an encapsulated motor-control system, but can be accessed when the location, orientation, and size of a particular object must be remembered. Such information would be useful, for example, when attempting to interact with objects in a familiar room or cave after the lights have been turned out.

A test for the neural locus of metric attribute storage would be to have patients with posterior parietal damage perform the episodic memory task for shape (Jolicoeur, 1987; Biederman & Cooper, 1992), described earlier, in which metric attributes (such as size) of the to-be-remembered shapes can vary from study to test. If the posterior parietal area does indeed store metric information, then patients with posterior parietal damage should reveal *less* interference for a change in a metric attribute between initial presentation and the subsequent episodic judgement of its shape.

This research was supported by AFOSR Research Grants 88-0231 and 90-0274 to I.B. and NSF Graduate Fellowships to E.E.C. and J.E.H. J.E.H. was also supported by an AFOSR Postdoctoral Fellowship on Research Grant 90-0274. We thank Peter C. Gerhardstein, Pierre Jolicoeur, and Keith Humphrey for careful readings of an earlier version of this manuscript. John Hummel is now at the University of California, Los Angeles. Correspondence concerning this article should be addressed to Irving Biederman, who is now at the Department of Psychology, University of Southern California, Hedco Neuroscience Building, University Park, Los Angeles, CA 90089-2520, E-mail address: ib@rana.usc.edu.

References

- Atkinson, R.C., & Juola, J.F. (1974). Search and decision processes in recognition. In D.H. Krantz, R. C. Atkinson, & P. Suppes (Eds.), *Contemporary Developments in Mathematical Psychology*, (Vol. 1, pp. 242-293). San Francisco: Freeman.
- Bartram, D. (1974). The role of visual and semantic codes in object naming. *Cognitive Psychology*, 6, 325-356.
- Besner, D., & Coltheart, M. (1975). Mental size scaling examined. *Memory & Cognition*, 4, 525-531.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94, 115-147.
- Biederman, I., & Cooper, E.E. (1991a). Priming contour deleted images: Evidence for intermediate representations in visual object priming. *Cognitive Psychology*, 23, 393-419.
- Biederman, I., & Cooper, E.E. (1991b). Object recognition and laterality: Null effects. *Neuropsychologia*, 29, 685-694.

- Biederman, I., & Cooper, E.E. (1991c). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20, 585-593.
- Biederman, I., & Cooper, E.E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 121-133.
- Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, 20, 38-64.
- Brooks, R.A. (1981). Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence*, 17, 205-244.
- Brody, B.A., & Pribram, K.H., (1978). The role of frontal and parietal cortex in cognitive processing: Tests of spatial and sequence functions. *Brain*, 101, 607-633.
- Bulthoff, H.H., Edelman, S., & Sklar, E. (1991). Mapping the generalization space in object recognition. Paper presented at the meeting of the Association for Research in Vision and Ophthalmology. Sarasota, FL.
- Bundesen, C., & Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Perception & Performance*, 1, 214-220.
- Bundesen, C., Larsen, A., & Farrell, J.E. (1981). Mental transformations of size and orientation. In J. Long, & A. Baddely (Eds.), *Attention and Performance IX*, (pp. 279-294). Hillsdale, NJ: Erlbaum.
- Cooper, E.E. & Biederman, I. (1992a). Congruency effects for position, orientation, and image features. Manuscript in preparation.
- Cooper, E.E., & Biederman, I. (1992b). *Priming objects with single volumes*. Manuscript submitted for publication.
- Ellis, R., & Allport, D.A. (1986). Multiple levels of representation for visual objects: A behavioural study (pp. 245-257). In A.G. Cohen & J.R. Thomas (Eds.) *Artificial Intelligence and its Applications*. New York: Wiley.
- Gerhardstein, P.C., & Biederman, I. (1991). Priming depth-rotated object images: Evidence for 3D invariance. Paper presented at the meeting of the Association for Research in Vision and Ophthalmology. Sarasota, FL. May.
- Goodale, M.A., Milner, D.A., Jakobson, L.S., & Carey, D.P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, 349, 154-156.
- Howard, J.H., & Kerst, S.M. (1978). Directional effects of size change on the comparison of visual shapes. *American Journal of Psychology*, 91, 491-499.
- Hummel, J.E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, in press.
- Jolicoeur, P. (1987). A size-congruency effect in memory for visual shape. *Memory & Cognition*, 15, 531-543.
- Jolicoeur, P., & Besner, D. (1987). Additivity and interaction between size ratio and response category in the comparison of size-discrepant shapes. *Journal of Experimental Psychology: Human Perception & Performance*, 13, 478-487.
- Jolicoeur, P., Gluck, M.A., & Kosslyn, S.M. (1984). Picture and names: Making the connection. *Cognitive Psychology*, 16, 243-275.

- Larsen, A. (1985). Pattern matching: Effects of size ratio, angular difference in orientation, and familiarity. *Perception & Psychophysics*, *38*, 63-68.
- Larsen, A., & Bundesen, C. (1978). Size scaling in human pattern recognition. *Journal of Experimental Psychology: Human Perception & Performance*, *4*, 1-20.
- Lowe, D.G. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision*, *1*, 57-72.
- Marr, D., & Nishihara, H.K. (1978). Representation and recognition of three dimensional shapes. *Proceedings of the Royal Society of London, Series B*, *200*, 269-294.
- Mishkin, M., & Appenzeller, T. (1987). The anatomy of memory. *Scientific American*, *256*, 80-89.
- Perenin, M.-T., & Vighetto, A. (1988). Optic ataxia: A specific disruption of visuomotor mechanisms. *Brain*, *111*, 643-674.
- Palmer, S.E. (1975). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In D.A. Norman & D. E. Rumelhart (Eds.), *Explorations in Cognition* (pp. 279-307). San Francisco: Freeman.
- Palmer, S.E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, *9*, 441-474.
- Pinker, S. (1984) Visual cognition: An introduction. In S. Pinker (Ed.), *Visual Cognition* (pp. 1-62). Amsterdam: Elsevier.
- Pohl, W. (1973). Dissociation of spatial discrimination deficits following frontal and parietal lesions in monkeys. *Journal of Comparative Physiological Psychology*, *82*, 227-239.
- Rock, I., & DiVita, J. (1987). A case of viewer-centered perception. *Cognitive Psychology*, *19*, 280-293.
- Shepard, R.N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701-703.
- Tarr, M.J. (1989). Orientation dependence in three-dimensional object recognition. Unpublished doctoral dissertation, Department of Brain and Cognitive Sciences, MIT.
- Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*, 169-193.
- Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*, 193-254.
- Ungerleider, L.G., & Brody, B.A. (1977). Extrapersonal spatial orientation: The role of posterior parietal, anterior frontal, and inferotemporal cortex. *Experimental Neurology*, *56*, 265-280.
- Ungerleider, L.G., & Mishkin, M. (1982). Two cortical visual systems. In D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield (Eds.) *Analysis of Visual Behavior* (pp. 549-586) Cambridge, MA: MIT.