Three-Dimensional Object Recognition is Viewpoint-Dependent

Michael J. Tarr, *Brown University*
Pepper Williams, *University of Massachusetts, Boston*
William G. Hayward, *University of Wollongong*
Isabel Gauthier, *Yale University*[*]

March 18, 1998

A fundamental assumption of many structural description-based theories of visual object recognition [1] is that the basic building blocks (primitives) of the descriptions should be recognized equally easily from any vantage point. We tested this assumption for the most prominent structural-description theory, Recognition-by-Components (RBC) [2], and its primitives. Counter to RBC's predictions, the results of nine separate experiments all revealed unequivocal viewpoint effects: Recognition of single primitives was progressively more difficult as the difference between studied and tested viewpoints increased. These findings call the validity of RBC into question and support theories positing view-based representations and recognition processes.

The human visual system is faced with the computationally difficult problem of recognizing objects in a three-dimensional (3D) world via two-dimensional (2D) retinal images. A widely accepted class of theories holds that to perform this task, humans first reconstruct 3D representations (structural descriptions) of to-be-recognized objects from the 2D retinal images, then match these representations to encoded structural descriptions. The seminal version of this approach proposed by Marr and Nishihara [1] was eventually superseded by Biederman's [2] more computationally-practical RBC theory, in which structural descriptions are assembled from sets of generalized cones ("geons"). Each geon is defined by a small set of visual properties that are present in the object's 2D projection when the geon is viewed from almost any position. For example, all three views of the brick in Figure 1 include a straight main axis, parallel edges, and a straight cross section. Objects are in turn defined by their component geons and the spatial relations in which the geons are placed—a mug is represented as a noodle side-attached to a cylinder, and a suitcase as a noodle top-attached to a brick. RBC's elegance and simplicity have led it to become the best-known theory of object recognition, described in almost every textbook on perception, cognition, and neuropsychology.

Since the defining properties of geons are almost always recoverable from images, a fundamental assumption of RBC is that recognition of geons, and by extension objects composed of geons, should be equally accurate and fast when seen from any viewpoint [6]. Do humans actually recognize objects in such a viewpoint-invariant manner? Clearly, recognition performance often *approaches* invariance, in that we rarely misidentify familiar objects despite variations in viewpoint produced by movement of the objects, ourselves, or both. In the psychophysics laboratory, viewpoint changes are simulated by having participants study pictures of objects taken from one viewpoint, then testing their recognition performance on pictures of the objects from the same or from different viewpoints. A number of studies [3, 4, 5] have demonstrated slower and/or less accurate recognition responses when different viewpoints were studied and tested, compared to when study and test views were identical. To account for these departures from RBC's predictions, Biederman and Gerhardstein [6] proposed that three "conditions for invariance" were violated in past demonstrations of viewpoint dependence: First, tested objects must be decomposable into geons; second, the structural descriptions of test objects must be qualitatively different from each other; and third, different views of each object must be parsable into the same structural description.

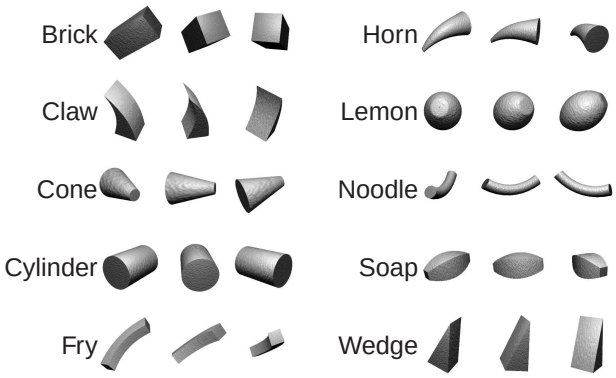

Figure 1: Shaded images of the three views of the ten geons used in the experiments, along with names assigned in Experiment 3. The leftmost figure in each row was arbitrarily designated the 0° view; the other two figures represent 45° and 90° rotations of the objects in the depth plane.

The imposition of such conditions obviously weakens the RBC theory, in that a separate mechanism will be needed to recognize classes of objects, such as faces, that do not meet the conditions for invariance. The question remains, however, whether viewpoint invariance is ob-

[*]This research was supported by an Air Force Office of Scientific Research Grant. We thank Jay Servidea and James Rosoff for their assistance in running the psychophysical studies. Address comments to Michael J. Tarr, Department of Cognitive and Linguistic Sciences, Brown University, Box 1978, Providence, RI 02912, email: Michael_Tarr@brown.edu. Version 1.2.4.

tained even when the conditions are met. In the present set of experiments, we assessed participants' ability to recognize single geons (Figure 1) when study and test images were identical or varied by 45°or 90° rotations in depth. Since geons are the basic primitives of RBC, this stimulus set obviously meets the conditions for invariance, so the theory *must* predict that they will be equally identifiable from any of the three viewpoints [6].

This point bears repeating: Although viewpoint-dependent recognition has been demonstrated in previous studies [3, 4, 5], the sets of stimuli used in these studies could be considered atypical in comparison to real-world objects, and thus not relevant for testing a theory of everyday object recognition such as RBC. The present study is not subject to this criticism, because the stimuli we used are the very building blocks of the theory in question. Recently-published reports [6, 7] including single experiments using single geons have produced discrepant results. Our set of experiments was designed as a definitive test of RBC's postulate that geons should be recognized in a viewpoint-invariant manner, including 9 experiments utilizing three different tasks (sequential matching, match-to-sample, and naming), two different versions of geons (line drawings and shaded images), and several other factors that might be expected to influence performance patterns.

Experiments 1A-1E utilized a sequential matching task, in which two images were presented back-to-back and participants decided whether they depicted the same or different geons (trials in which different geons were presented were not of theoretical interest, so only the results of "same" trials are discussed). The null hypothesis, that geons were recognized without a cost for changes in viewpoint, is unsupportable for any of the five experiments (see Figure 2 for results of all 9 experiments). The same conclusion holds for results of Experiments 2A-2C, which used a match-to-sample task. The apparent reduction in the size of viewpoint effects here compared to Experiments 1A-1E was at least in part a result of practice effects. In the match-to-sample procedure, trials were run in blocks, where a participant saw a target geon in the 0° view, followed by three trials each of the 0°, 45°, and 90° views of the target geon, interspersed with 9 other geons. For the first trials in each block, the difference between 0° and 90° views averaged 52 ms (averaged over all 85 participants in Experiments 2A-2C), whereas for the third trials with each view, the difference averaged only 22 ms. This interaction of trial number and viewpoint was significant, $F(2, 336) = 4.62$, $p = .0012$, and is reminiscent of results from other experiments [8, 9, 5] in which viewpoint effects were stronger in initial than in later blocks of trials. Indeed, the same type of practice effect was observed for the naming task used in Experiment 3, in which participants first learned labels for the 0° view of each geon, then were asked to name 0°, 45°, and 90° views in two subsequent blocks. The effect of viewpoint difference was significant in the first block, but not in the second, and the interaction of block and viewpoint difference was again significant, $F(2, 48) = 7.51$, $p = .0015$.

Although viewpoint effects in each experiment were significant, it is possible that these overall patterns were the result of a small subset of anomalous participants and/or stimulus items. We tested for this contingency with non-parametric sign tests, which indicated that across the 9 experiments, 86% of participants and an average of 9.3 of the 10 geons were faster for 0° viewpoint changes than 90° viewpoint changes, 76% of participants and 8.2 geons were faster for 0° viewpoint changes than 45° viewpoint changes, and 70% of participants and 7.3 geons were faster for 45° viewpoint changes than 90°viewpoint changes, all z's > 4.43, all p's $< .001$.

One other aspect of our data should be pointed out here. In the match-to-sample and naming experiments (Experiments 2A, 2B, 2C, and 3), the 0° view was always the "studied" view and participants were tested on 0°, 45°, and 90° views. In the sequential matching procedure of Experiments 1A-1E, however, all pairwise combinations of views were tested. That is, the 0° viewpoint difference condition includes trials testing all three of the 0°, 45°, and 90° views in Figure 1, the 45° viewpoint difference condition includes trials testing 0°-45° view combinations and 45°-90° view combinations (in both possible orders), and the 90° viewpoint difference condition includes 0°-90° trials and 90°-0° trials. Thus the viewpoint-dependent patterns revealed in these experiments could not have been due to certain views being inherently easier or harder to process and/or remember. The decrease in performance from 0° to 45° to 90° conditions results from the viewpoint *changes* in the latter two conditions, not from the particular views that were tested.

RBC holds that visual recognition is performed by describing the 3D structure of objects in terms of geons. Since geons are specifically designed to be equivalently recognizable from almost any viewpoint, RBC predicts that object recognition should be viewpoint-invariant (as long as the conditions specified in [6] are satisfied). The present results invalidate this claim: In all 9 experiments, geon recognition was viewpoint-dependent. Clearly, structural descriptions based on geons cannot be recovered in a viewpoint-invariant manner if recognition of geons themselves is viewpoint-dependent.

These findings are at odds with those of Experiment 4 of Biederman and Gerhardstein's study [6], in which the procedure was very similar to our Experiment 2C but only small viewpoint effects were found. However, Biederman and Gerhardstein employed several experimental conditions that appear especially well-suited to obtaining viewpoint invariance: The combination of match-to-sample task, go/no-go procedure, and high-contrast line drawings may have allowed participants to focus on one or two diagnostic features of the initially presented target, an unrealistic strategy for object recognition in most other situations. Furthermore, as was evident in our data, practice effects in the match-to-sample task sharply reduce the size of viewpoint effects when results are averaged over initial and subsequent trials within blocks (trial order effects were not reported in [6]).

Could RBC somehow be reconciled with the present results? Given the wide variety of procedural variations used here, along with the fact that the stimuli employed were the very atoms of RBC's representational system, explaining the findings away with new "conditions" [6] does not seem a viable option. A second possible response, that the viewpoint effects found here were relatively small and thus inconsequential, is also untenable: When considered as a percentage of baseline response times (that is, response times for 0° conditions, in which no viewpoint change occurred), 90° viewpoint changes led to recognition time costs of between 6.2% (Experiment 2B) and 12.9% (Experiment 1E), non-trivial effects that must be explained by any comprehensive theory of human object recognition.

In fact, these substantial and robust viewpoint effects

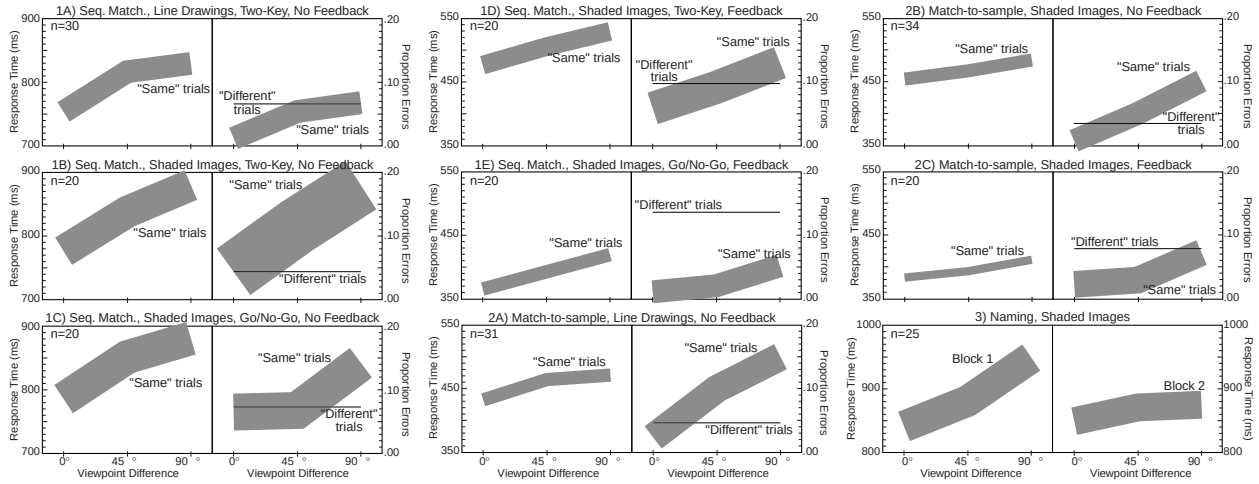

Figure 2: Results from sequential matching (1A-1E), match-to-sample (2A-2C), and naming (3) experiments. Tasks, major procedural differences, and numbers of participants are given for each set of graphs. The width of all shaded lines are proportional to two times the within-participants 95% confidence intervals for each experiment [19]. Since perfectly flat functions would not fit within any of these shaded areas (with the exception of Block 2 of Experiment 3), the viewpoint-invarant pattern of effects can be safely rejected in every experiment. F values for ANOVAs on the data from each experiment ranged from 8.38 and 29.08 for response times and from 5.22 and 32.58 for error rates, all significant at the $p < .01$ level.

for extremely simple 3D volumes argue against any theory proposing viewpoint-invariant representations. Furthermore, our findings are convergent with neurophysiological studies of temporal and parietal lobe structures thought to be responsible for object identification performance [10]. For example, neurons responsive to human faces have been found in the Macaque superior temporal sulcus (STS). The majority of these cells were also preferential for specific views of faces [12]: If a particular cell responded best to a frontal view, rotating the face 45° in depth or in the picture-plane dramatically reduced the activation of that neuron, and a 90° rotation almost completely extinguished the cell's response. Similar "view-tuned" cells have been found that show the greatest activation to face profiles, backs of heads, lowered heads, or raised heads. Comparable results have also been obtained with novel computer-generated 3D objects ("paperclips" and "spheroids," [11]) that monkeys were trained to recognize. Following training, recordings in inferior temporal cortex (IT) revealed neurons that responded preferentially to previously unfamiliar objects, and as with faces, different cells were found to be sensitive to different views of the same object.

Taken together, the results presented in this paper, the results of these neurophysiological studies, and the results of previous behavioral studies offer persuasive evidence that object recognition is an inherently viewpoint-dependent process. This perspective is embodied in a host of theories [13, 14, 15, 16] which assume that collections of features, surfaces, parts, or entire images of objects are represented in a viewpoint-specific manner. According to these theories, recognition of test objects is based on the similarity between studied images and tested images. Objects seen from viewpoints increasingly different from learned views will project increasingly less-similar images, so view-based theories provide a natural account for viewpoint effects found here and in other recent recognition studies [7, 17].

Methods

220 Yale University undergraduates participated in exchange for course credit or cash payment; numbers of participants in each individual experiment are given in Figure 2. Line drawings of three viewpoints of ten single geons were scanned into a Macintosh computer from Biederman and Gerhardstein's [6] Figure 12 for use in Experiments 1A and 2A. Shaded images of the same 10 geons (Figure 1), matching the views used by Biederman and Gerhardstein as closely as possible, were created and rendered using CAD software for use in all other experiments. All stimuli subtended approximately 7° by 7° of visual angle when viewed by participants approximately 60 cm from the computer screen. All experiments were performed on Macintosh computers using RSVP software (http://psych.umb.edu/rsvp).

Each sequential matching trial (Experiments 1A-1E) consisted of the following sequence of events: blank screen for 1000 ms, fixation cross for 500 ms, object image for 200 ms, mask (consisting of random combinations of features from the line drawings or shaded images) for 750 ms, second object image for 100 ms, mask for 500 ms. The trial timed out 1500 ms later if no response was given. In Experiments 1A, 1B, and 1D, a two-key procedure was used, in which the participant pressed the "V" key on the computer keyboard if the two images were of the same geon (even if shown in a different viewpoint), or the "M" key if the two images were of different geons. In Experiments 1C and 1E, a go/no-go procedure was used, in which the participant pressed the space bar if the two objects were the same or allowed the trial to time out otherwise. Participants in Experiments 1D and 1E were informed after each trial of their response time and the accuracy of their response (as seen in Figure 2, this feedback lowered overall response times, but had little impact on viewpoint effects). For "same" trials, each of the three views of each geon was presented three times as the first object in a trial and three times as the second object, producing three trials in which the two views were identical, four trials in which they differed by 45°, and two trials in which they differed by 90°. In these and subsequent experiments, trials were presented in a different random order for each participant.

Match-to-sample experiments (2A-2C) each consisted of 10 blocks of trials. Each block included an initial presentation of a target object for 20 s, followed by 18 test trials which followed the sequence: blank screen for 250 ms, fixation cross for 500 ms, object for 150 ms, mask for 500 ms, time out 1500 ms later if no response was given. Participants memorized the initial target, then pressed the space bar if the test object on subsequent trials matched the target object (even if the viewpoint varied), or let the trial time out if the test and target objects did not match. Participants in Experiment 2C received feedback of the same sort as Experiments 1D and 1E. The target object was always shown in the 0° viewpoint. Every test block included three "same" trials in each of the three viewpoints (0°, 45°, and 90°) of the target object and nine "different" trials (one each of the non-target objects).

In Experiment 3, participants learned labels (given in Figure 1) for the 10 geons, then performed trials in which they verbally named test images as quickly as possible. Participants first studied a sheet of paper showing the 10 objects with their names, then performed 20 trials in which each object was shown twice, along with its name, on the computer screen. Four practice blocks of five trials with each object followed, in which participants saw objects without their names and spoke the names. Objects were always shown in the 0° viewpoint during practice trials. Participants then performed two test blocks of six trials with each object, distributed equally between the 0°, 45°, and 90° viewpoints. All practice and test trials consisted of a 500 ms blank screen, 500 ms fixation cross, and an object that stayed on the screen until the participant responded or until 2500 ms had elapsed. Response times were recorded via the voice trigger, but accuracy was not recorded (the experimenter observed the first few participants, and found that accuracy was almost always perfect).

References

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

[2] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[3] H. H. Bülthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Acadamy of Sciences USA*, 89:60–64, 1992.

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

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

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

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

[8] P. Jolicoeur. The time to name disoriented natural objects. *Memory and Cognition*, 13:289–303, 1985.

[9] M. J. Tarr and S. Pinker. Mental rotation and orientation dependence in shape recognition. *Cognitive Psychology*, 21:233–282, 1989.

[10] M. S. Cohen, S. M. Kosslyn, H. C. Breiter, G. J. DiGirolamo, W. L. Thompson, A. K. Anderson, S. Y. Bookheimer, B. R. Rosen, and J. W. Belliveau. Changes in cortical activity during mental rotation: A mapping study using functional MRI. *Brain*, 119:89–100, 1996.

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

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

[13] E. Bricolo, T. Poggio, and N. K. Logothetis. 3D object recognition: A model of view-tuned neurons. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 41–47. MIT Press, Camrbidge, MA, 1997.

[14] S. Edelman and D. Weinshall. A self-organizing multiple-view representation of 3D objects. *Biological Cybernetics*, 64:209–219, 1991.

[15] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.

[16] D. I. Perrett, M. W. Oram, and E. Wachsmuth. Evidence accumulation in cell populations responsive to faces: An account of generalisation of recognition without mental transformations. *Cognition*, in press.

[17] M. J. Tarr, H. H. Bülthoff, M. Zabinski, and V. Blanz. To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science*, 8(4):282–289, 1997.

[18] R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

[19] G. R. Loftus and M. E. J. Masson. Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review*, 1:476–490, 1994.