PERGAMON

Vision Research 40 (2000) 3895–3899

Vision Research

www.elsevier.com/locate/visres

Differing views on views: comments on Biederman and Bar (1999)

William G. Hayward [a,*], Michael J. Tarr [b,1]

[a] *Department of Psychology, Chinese University of Hong Kong, Shatin, NT, Hong Kong*
[b] *Department of Cognitive and Linguistic Sciences, Brown University, Box 1978, Providence, RI 02912, USA*

Received 20 October 1999; received in revised form 25 May 2000

In a recent article, Biederman and Bar (1999) present several results to support "a class of theories [that] assumes that non-accidental properties (NAPS) might be exploited so that even novel objects can be recognized under depth rotation" — specifically, theories based on 'geons' (Biederman, 1987; Hummel & Biederman, 1992). Biederman and Bar likewise present results that they believe to be inconsistent with a "class of theories … based on generalization from templates specified by metric properties" — specifically, 'view-based' or 'image-based' theories (Bricolo, Poggio, & Logothetis, 1997; Poggio & Edelman, 1990; Tarr & Bülthoff, 1995, 1998). Because our disagreements with Biederman's theoretical approach have been detailed in many other forums (Hayward & Tarr, 1997; Tarr & Bülthoff, 1995, 1998) we do not feel it is necessary to respond in kind. We do feel, however, that it is necessary to address a series of straightforwardly incorrect claims made by Biederman and Bar regarding our published results. Below we enumerate these claims, and our replies to each[2]:

1. "*A task in which subjects are trained with arbitrary names for a particular pose of an object (Tarr, Williams, Hayward, & Gauthier, 1998, Experiment 3; Haywood (sic) & Tarr, 1997, Experiment 2), particularly if the distinguishing information is difficult to discriminate, is problematic insofar as pose is part of what is learned and, potentially, used" (p. 2895).*

According to this statement, our cited experiments produced results that are artifactual because names for objects were learned only at specific viewpoints. There are at least three reasons to conclude that no such problems exist. First, in these and other studies, we obtained the same pattern of viewpoint dependency with identification (naming) tasks *and* with Biederman and Bar's preferred same–different matching tasks (Hayward & Tarr, 1997; Tarr, Bülthoff, Zabinski, & Blanz, 1997; Tarr et al., 1998). Moreover, we almost always employed a same–different matching procedure that was identical to (and intentionally based on) that used by Biederman and Gerhardstein (1993).

Second, in these same three studies (Hayward & Tarr, 1997; Tarr et al., 1997; Tarr et al., 1998) we explicitly used stimuli ('geons' and 'geon objects') which were extremely easy to discriminate. Indeed, most stimuli were almost identical to those used by Biederman and Gerhardstein (1993). These studies show the same pattern of responses as studies using highly similar objects that are difficult to discriminate (e.g. Tarr, 1995; Tarr & Pinker, 1989). Thus, there is relatively little evidence, as suggested by Biederman and Bar, that recognition of highly similar objects is qualitatively different from recognition of geons (for additional evidence, see Hayward & Williams, in press).

Third, Biederman and Bar appear to be intimating that pose should be divorced from training on novel objects. To the extent that this is possible, it has certainly been accomplished in several studies, either by presenting animated objects rotating back and forth in depth (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992), by training subjects with the standard and mirror-reversed versions of each object (so that subjects would never have to do the equivalent of distinguishing a *p* from a *q* or a *b* from a *d*) (Tarr, 1995; Tarr & Pinker, 1989), or by training subjects with multiple views of each object (Tarr, 1995; Tarr & Gauthier, 1998; Tarr & Pinker, 1989). Such manipulations would

* Corresponding author. Tel.: +852-26096195; fax: +852-26035019.

E-mail addresses: william-hayward@cuhk.edu.hk (W.G. Hayward), michael_tarr@brown.edu (M.J. Tarr).

[1] Comments can also be addressed to MJT. Tel.: +1-401-8631148; fax: +1-401-8632255.

[2] Because some of the points raised in this letter relate directly to the appearance of stimuli we have used in various experiments, readers are encouraged to examine our stimuli for themselves — all of which are available for download at our web site: http://www.cog.brown.edu/~tarr.

PII: S0042-6989(00)00179-6

seem to dissociate shape and orientation as experimental confounds, yet results of all these studies show clear viewpoint dependence. Thus, any concatenation of object shape and viewpoint in our studies must almost certainly be a consequence of the nature of mental representations of objects, not idiosyncratic training conditions.

2. "*… many of the views in Tarr et al. (1998) were, in fact, near accidents that required, for example, determination of whether a single small contour was straight or slightly curved to distinguish one object from another*" (pp. 2895–6).

In responding to this point, it is important to reiterate that we used objects (geons) that were identical to those used in Biederman's earlier work *and* that we matched the specific viewpoints of each object to the viewpoints used by Biederman and Gerhardstein (1993). Consider that in one instance (Tarr et al., 1998; Experiments 1a and 2a) we went so far as to digitally scan the actual objects and viewpoints illustrated in their paper and simply used these identical images in our study.

Moreover, the observed data do not support Biederman and Bar's conclusion. As evidence, they cite differences in the false alarm rates in comparable tasks from their earlier study (Biederman & Gerhardstein, 1993) and our study (Tarr et al., 1998). First, the high false alarm rates reported by Biederman and Gerhardstein (1993) (sometimes as high as 60–100 and 15–20% overall) indicate that subjects had relatively low sensitivity in discriminating targets from distractors. In contrast, our subjects had relatively higher sensitivity for the same discriminations. Since we presume the goal of any theory of recognition is *accurate* recognition, it seems more likely that the poor performance obtained by Biederman and Gerhardstein does *not* reflect standard recognition processes.

3. "*Before considering the details of how confounds of resolution and other factors might have artifactually produced rotation costs in these other studies, we note that there is independent evidence that resolution variations may be sufficient to produce the observed rotation costs. Curiously, most such experiments reporting significant slopes with (presumably) distinctive NAPs have studied relatively small rotation angles, up to 90° and, in some cases, only to about 30° (Haywood (sic) & Tarr, 1997)*" (p. 2896).

This statement implies that the use of small rotation angles (as opposed to a 180° range) in some of our studies restricts our tests of viewpoint invariance to conditions where we were more likely to find viewpoint costs. Such an interpretation is unwarranted.

First, small rotations provide *the strongest test* of the two theories in question. Across large rotations of an object, changes in its image are almost always far more dramatic, including changes in visible surfaces, changes in configuration, and, perhaps most importantly, changes in NAPs (Tarr & Kriegman, 1999). Thus, large rotations are highly likely to produce performance costs, but these are often uninterpretable in that they may be attributed in a theory-appropriate way to either viewpoint-dependent normalization mechanisms or to mismatches between different structural descriptions. Indeed, Biederman and Gerhardstein (1993) heavily emphasize this point — making it clear that studies that test view invariance across changes in NAPs do not provide data that bear on whether human recognition is view invariant as per Biederman's stated theory. In contrast, small rotations allow us to test whether visual recognition is indeed viewpoint dependent even in instances where relatively little visual information changes between rotations. For example, in our experiments using rotations only up to 30° (Hayward & Tarr, 1997), there were *no changes to the visible parts* of any of the objects shown in the study. We believed that this manipulation allowed us to compare the predictions of view-based theories with those of Biederman's theory. The combination of Biederman and Bar's suggestion that such small rotations are inappropriate *and* Biederman and Gerhardstein's (1993, 1995) observation that larger rotations suffer from changes in visible parts results in an extremely restricted range of conditions for which Biederman's theory still holds: those that are large enough to remove so-called 'resolution confounds', but small enough that they do not change which parts are visible.

Second, just as the literature indicates that viewpoint costs do not depend upon the type of stimuli used in an experiment (point #1), it also shows costs across precisely the range of viewpoints advocated by Biederman and Bar. For example, the range of viewpoints used in Tarr et al. (1998) spanned 180°; 90° in each direction from the training viewpoint. Moreover, Hayward (1998) found viewpoint costs in a recognition study across a 180° rotation. Many studies in other laboratories have found similar results across large rotations (Lawson & Humphreys, 1996; Newell & Findlay, 1997; Srinivas, 1995).

Finally, we object to Biederman and Bar's characterization of the shape differences used in our studies as 'presumably' NAP differences. Even those objects (e.g. from Hayward & Tarr, 1997) that were not directly based on stimuli used in Biederman and Gerhardstein (1993) have clear NAP differences as defined by Biederman (1987). In addition, all our stimuli are available for inspection in our papers and on the web.

4. "*From a view-based perspective, a rotation of 180° or mirror-reflection of a bilaterally symmetrical object would be expected to produce enormous rotation costs, relative to these slight rotation angles*" (p. 2896).

To make this assumption, Biederman and Bar adopt an interpretation of view-based theories as metric 'tem-

plate' theories, a point made frequently in their paper. Even many of the earliest instantiations of view-based models predict mirror-reflection invariance for objects that show the *same* features in the front and back (e.g. Poggio & Edelman, 1990). More recent models make this prediction much more explicit, specifically modeling the non-linearities that occur at 180° in physiological data recording from monkeys (Bricolo et al., 1997; Riesenhuber & Poggio, 1999). Moreover, these models *are not* metric templates in the simplistic sense that Biederman and Bar suggest. For instance, they often employ a hierarchy of image-based features to represent and recognize objects. As such they exhibit invariance over mirror-reflections because the same local image-based features will be present in both the standard and mirror versions of an image. Similarly, models using image-based features are invariant across many of the image changes that arise from configural deformations, and changes in translation and scale.

Moreover, the recognition of objects across mirror-reflections *is not* the same as the recognition of objects across 180° rotations in depth. If the objects in question have different features visible in their front and back views (as with almost all objects other than those that are radially symmetric), then a 180° rotation changes the actual visible structure of the image, not only the parity of the object's contours. For example, Biederman's studies on this issue have often used a 3/4's view of an airplane either pointing to the left or to the right; in contrast, a 180° rotation from one of these views would reveal features on the *back* of the airplane that are unseen from the initial view. Thus, the data cited by Biederman and Bar (Biederman & Cooper, 1991a,b; Stankiewicz, Hummel, & Cooper, 1998) *do not* bear on the issue of whether human recognition performance is invariant for the 180° case.

5. *"The procedures of the present investigation were designed to minimize the artifacts that can lead to apparent rotation costs with stimuli that differ in NAPs. The essential point here is that rotation in depth tends to produce drastic changes in the 2D image that can differentially affect the perceptibility of the parts"* (p. 2896).

This statement implies that earlier studies obtained viewpoint costs because the perceptibility of the stimulus images differed across rotations. First, the critical finding is not simply that depth rotations hinder recognition performance. Rather, in many studies we (and others) have obtained a *systematic* (and usually monotonic) increase in recognition costs as objects are rotated further from trained or known views (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Gauthier & Tarr, 1997; Lawson, Humphreys, & Watson, 1994; Tarr, 1995; Tarr et al., 1997; Tarr & Gauthier, 1998; Tarr & Pinker, 1989, 1990; Tarr et al., 1998). For this systematic pattern to be an artifact, it would have to be the case that larger rotations *progressively* produced greater disturbances to the *perceptibility* of the stimuli (not simply in the difference between the known image and the image following rotation).

Second, in Hayward and Tarr (1997), Tarr et al. (1997) and Tarr et al. (1998), we employed a sequential-matching task in which we rendered many views for each object and generated equal magnitude rotations by using pairs of views separated by the appropriate depth rotation. Thus, data on a rotation of particular magnitude are the result of many different pairs of views. Given that view pairs from such an equivalence class consistently produced highly similar performance, it seems unlikely that differences in 'perceptibility' *just happen* to be equal for each such pair in an equivalence class. Moreover, one of the critical predictions of view-based models, a systematic increase in recognition costs with increasing separation in depth rotation, was once again found when these equivalence classes were compared to one another. That is, equal magnitude view pairs consistently yielded better performance when the depth separation was small and progressively poorer performance as the depth separation increased. It is therefore essentially impossible that on large rotations one or both of the stimuli were difficult to perceive, but on small rotations they were easy to perceive. Rather, across large rotations observers had greater difficulty determining that the two images represented different views of the same object.

6. *"Rendered images, as compared to line drawings, typically have lower contrast and illumination and shadow contours that can increase the difficulty of determining the orientation and depth discontinuities important for resolving the geons. Such resolution difficulties characterize the object images in the Tarr et al. (1997) and Haywood (sic) and Tarr (1997) experiments"* (p. 2896).

This statement clearly indicates that Biederman and Bar believe that rendered images (e.g. photorealistically shaded images of 3D objects) are somehow less appropriate stimuli than line drawings for studying human visual recognition. First, we strongly argue against the notion that line drawings provide a more suitable stimulus domain than rendered images. Not only is it trivially obvious that 'real-world' recognition *always* involves recognition of textured and shaded objects against richly textured and shaded backgrounds, but it is not even clear that the human visual system can extract the kind of line drawing that Biederman and his collaborators have used in their studies (Kurbat, 1994; Sanocki, Bowyer, Heath, & Sarkar, 1998).

Second, in all of our studies using rendered images, the stimuli contained regions of high contrast around the outlines of the objects, and these outlines often provided sufficient NAP differences to perform the recognition task (see also Hayward, 1998). Thus, even if rendered images can be made difficult to perceive by

intentionally embedding them in cluttered scenes, we never used such manipulations to idiosyncratically enhance the potential effects of depth rotations.

Third, this point ignores at least two published findings. Tarr et al. (1998) used rendered images *and* line drawings of the identical objects – the latter being scanned in directly from Biederman and Gerhardstein (1993). In that study we obtained nearly indistinguishable results for the rendered versions and line drawing versions of the stimuli. Likewise, Biederman and Ju (1988) compared recognition of line drawings and color photographs of objects. As in our study, they found no consistent differences in patterns of recognition for the two types of stimuli, and concluded that colored photographic images can be identified about as quickly as line drawings of the same objects.

7. *"Biederman and Bar argued that these transient shifts were the reason why, in the Haywood (sic) and Tarr (1997) and Tarr et al. (1997) studies, a rotation from 0° to a slight angle, say 30°, produced greater costs than rotations from greater angles, say from 60° to 90°. The opposite would be expected from the template extrapolation/mental rotation routines argued by Tarr (1995)"* (p. 2896).

We are unsure of the specific claim here. One interpretation is that we found viewpoint costs which were larger at 30° than at any other rotation. A second interpretation is that the difference between no rotation and 30° was larger than the difference between rotating 60° and rotating 90°. Both of these claims are incorrect.

In absolute terms, rotation costs in our studies almost always increase monotonically, so that costs at 60° are larger than costs at 30°, and costs at 90° are larger than those at 60°. Such is the case in the specific studies that Biederman and Bar cite: In Tarr et al. (1997) rotation costs were always greater for the 60° and the 90° rotation conditions as compared to the 30° condition. In Hayward and Tarr (1997) we never tested beyond 30°, but the 10° and 20° rotation conditions always produced smaller rotation costs. Likewise, in a third relevant study, Tarr et al. (1998) used rotations of 45° and 90° and across *nine* experiments and found that the 90° condition always produced larger rotation costs as compared to the 45° condition.

In relative terms, the rotation costs we have found in these three studies are almost always a function of the magnitude of the rotation difference, which results in any 30° difference in rotations, e.g. 0–30° or 60–90°, having a similar effect on performance. Critically view-based theories predict this pattern, that is, *equivalent* costs for *equal magnitude* rotations, *not* larger costs for the 60–90° case as compared to the 0–30° case as Biederman and Bar's statement would suggest ("The opposite would be expected …")[3].

In addition, we are skeptical regarding the existence of the 'transient shift' detectors which are hypothesized to detect a change to the image when there is no spatial translation between two sequentially presented stimuli. To our knowledge, there is no evidence that in a sequential-matching paradigm across a *masked*, 750 ms interstimulus interval, "IT cells have a transient response, originating in the magnocellular system, to any stimulus change" (p. 2887). Any magnocellular cell responses will be driven by the mask, and so it seems physiologically implausible that these neurons, 750 ms later, are able to revert to some pre-mask firing state to assist in the recognition judgment. In addition, there is behavioral evidence that a masked ISI as brief as 100 ms successfully eliminates image-based (or iconic) processing from recognition judgments (Ellis & Allport, 1986). Finally, the issue of whether transients are responsible for apparent rotation costs has been tested in recent experiments. Hayward and Williams (in press) conducted a sequential-matching task for objects that contained NAP differences, and, on each trial, presented the two stimuli in different spatial locations. This is precisely the manipulation advocated by Biederman and Bar in order to eliminate the transient. In this study, however, Hayward and Williams obtained exactly the same type of monotonically increasing viewpoint cost functions observed in our earlier work.

Acknowledgements

We thank two anonymous reviewers for their comments on this manuscript. MJT was supported by NSF award # SBR-9615819.

[3] One caveat is needed here. These viewpoint costs are always associated with changes in viewpoint, which have in the past resulted in explanations, by us and others (Jolicoeur, 1990; Tarr, 1995), that appeal to mental rotation processes. Recent theoretical formulations, however, have eschewed such mechanisms (Edelman & Weinshall, 1991; Perrett, Oram, & Ashbridge, 1998; Riesenhuber & Poggio, 1999; Tarr & Bülthoff, 1998). Viewpoint costs should, to our minds, be considered as due to changes in information in the relevant visual images. It appears that these changes in information are almost always closely associated with changes in viewpoint. However, there is no directly causal relationship between the magnitude of a change in viewpoint of an object, and the magnitude of the associated cost to recognition.

References

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

Biederman, I., & Cooper, E. E. (1991a). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, *20*, 585–593.

Biederman, I., & Cooper, E. E. (1991b). Priming contour-deleted images: evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, *23*, 393–419.

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

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

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*, 38–64.

Biederman, I., & Bar, M. (1999). One-shift viewpoint invariance in matching novel objects. *Vision Research*, *39*, 2885–2899.

Bricolo, E., Poggio, T., & Logothetis, N. K. (1997). 3D object recognition: a model of view-tuned neurons. In M. C. Mozer, M. I. Jordan, & T. Petsche, *Advances in neural information processing systems 9* (pp. 41–47). Cambridge, MA: MIT Press.

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 of the USA*, *89*, 60–64.

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.

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

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

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

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

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

Hayward, W. G., & Williams, P. Viewpoint dependence and object discriminability. *Psychological Science* (in press).

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

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

Kurbat, M. A. (1994). Structural description theories: is RBC/JIM a general-purpose theory of human entry-level object recognition? *Perception*, *23*, 1339–1368.

Lawson, R., & Humphreys, G. W. (1996). View specificity in object processing: evidence from picture matching. *Journal of Experimental Psychology*: *Human Perception and Performance*, *22*, 395–416.

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

Newell, F. N., & Findlay, J. M. (1997). The effect of depth rotation on object identification. *Perception*, *26*, 1231–1257.

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

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

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

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

Srinivas, K. (1995). Representation of rotated objects in explicit and implicit memory. *Journal of Experimental Psychology*: *Learning*, *Memory and Cognition*, *21*, 1019–1036.

Stankiewicz, B. J., Hummel, J. E., & Cooper, E. E. (1998). The role of attention in priming for left-right reflections of object images. *Journal of Experimental Psychology*: *Human Perception and Performance*, *24*, 732–744.

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

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

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

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

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

Tarr, M. J., & Kriegman, D. J. (1999). Toward understanding human object recognition: aspect graphs and view-based representations (submitted).

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

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

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