Max–Planck–Institut
für biologische Kybernetik
Spemannstraße 38 • 72076 Tübingen • Germany

Technical Report No. 54

Viewpoint Effects in Naming Silhouette and Shaded Images of Familiar Objects

Jeffrey C. Liter[1], Bosco S. Tjan[2],
Heinrich H. Bülthoff[3], & Nicole Köhnen

December 1997

Presented at the 38th Annual Meeting of the Psychonomic Society, Philadelphia, PA, November 20-23, 1997. This research was supported by a grant from the Human Frontier Science Program to J. Liter and by a Max-Planck Postdoctoral Stipend to B. Tjan.

[1] AG Bülthoff, E–mail: jeffrey.liter@tuebingen.mpg.de
[2] AG Bülthoff, E–mail: bosco.tjan@tuebingen.mpg.de
[3] AG Bülthoff, E–mail: heinrich.buelthoff@tuebingen.mpg.de

Viewpoint Effects in Naming Silhouette and Shaded Images of Familiar Objects

Jeffrey C. Liter, Bosco S. Tjan, Heinrich H. Bülthoff, & Nicole Köhnen

Abstract. We studied the visual features that support efficient entry-level object recognition by measuring naming latencies for different views of artifacts and four-legged animals that were shown as shaded images or as silhouettes. Experiment 1 revealed important differences in performance for the two renderings. Although three-quarter views of animals were recognized relatively quickly when shaded, they were not recognized quickly when presented as silhouettes. The same was true of artifacts when they were seen from the back. Experiment 2 used ideal-observer analyses to confirm that these effects could not be accounted for by differences in the intrinsic complexity of the stimuli. Together, these findings indicate that, for human observers, the shape of an object's bounding contour does not serve as a direct visual coding of the object, although it might be used as a first index into visual memory. These results also indicate that shading is important for recognizing objects in certain views. It remains unclear, however, what features are provided by shading. Shading might be used to derive a more precise part-based description of the object, or it might be used to extract surface properties or distinctive features and their spatial relations that are themselves elements of a view-based description of the object.

1 Introduction

Theoretical accounts of long-term visual memory for objects rely on different representations of object shape, including object-centered, three-dimensional (3-D) models (Marr & Nishihara, 1978), part-based structural descriptions (Biederman, 1987), and collections of viewer-centered, image-based descriptions (Bülthoff, Edelman, & Tarr, 1995; Perrett et al., 1984). Successful recognition according to these approaches will depend on the presence in the image of different features. For example, Marr (1982) argued that successful recognition using 3-D models depends on the ease with which the object's primary axis of elongation can be derived from the image, whereas Biederman and Gerhardstein (1993) argued that recognition depends on the visibility of an object's parts.

Because the presence of such features in the image necessarily depends on the direction in space from which an object is seen, it is not surprising that researchers have found differences in the ease with which different views of objects are recognized. Palmer, Rosch, and Chase (1981) found that the time needed to name familiar objects varied for different views of the objects, with subjectively preferred views named faster than nonpreferred views. These "canonical" views included profile views and three-quarter views (i.e., views seen from approximately 12-15 degrees of elevation from the ground plane and 30-45 degrees of rotation about the vertical axis from the full frontal view). These views were named quickly and accurately, whereas views from higher elevations and views from the back were named more slowly and with more errors.

Similar results were obtained in experiments carried out using 3-D computer graphic models of common objects. Blanz, Tarr, Bülthoff, and Vetter (1996) had observers select preferred views of objects by rotating the computer models of the objects in real time using a "Spaceball." As in Palmer et al.'s experiments, three-quarter and profile views were preferred over other views. Subsequent naming experiments carried out by Liter and Bülthoff (1997) confirmed that these preferred views were, in fact, named faster than nonpreferred views.

Which views of an object turn out to be canonical should, of course, depend on what visual features are important for the representation that is utilized. If the representation relies on axis-based, 3-D models, then views in which the primary axis of elongation is visible should be recognized more easily than views in which the axis is not visible. If the representation relies on part-based structural descriptions, then views in which important parts are occluded should be more difficult to recognize than views in which the important parts are visible.

The factors that make a view of an object canonical or noncanonical are as yet not entirely clear. Palmer et al. (1981) found that view canonicality correlates well with the visibility of surfaces that are rated as important for specifying the identity of the object. These views could also be characterized as those likely to be seen when interacting with the object. Thus, Palmer et al. concluded that familiarity and function are important in specifying view canonicality. This sort of explanation, however, helps little to constrain the nature of long-term visual representations. For example, these results could occur with an axis-based representation if the 3-D object models had an intrinsic orientation that matched the most familiar view of the object. Alternatively, these results could occur with a multiple-views representation if different views were weighted according to familiarity.

To gain insight into the visual representation itself, we need a better understanding of which visual features influence view canonicality. One way to approach this problem is to degrade views of objects that are known to be canonical and examine whether they are still easily recognized. In the present study we will investigate whether silhouette images of objects from familiar entry-level categories show the same pattern of canonical view effects as do shaded images. If the visual features that make a view canonical are based on or can be derived from the shape of an object's bounding contour, then silhouette images should show the same pattern of canonical view effects as shaded images.

Our interest in studying silhouette images is based partly on the success of computational recognition algorithms that utilize only the information contained in the bounding contour and on recent psychophysical results with silhouette images. Hayward (1995), for example, studied naming and sequential matching using both novel and familiar objects. In the matching experiments, observers viewed a shaded image of an object in a canonical view and then decided whether a subsequently presented image (either shaded or silhouette) showed the same object. The viewpoint was sometimes different in the two intervals. Response times were fast if the second image was shaded, but only if the same parts of the object were visible in both intervals. These findings were in agreement with those of Biederman and Gerhardstein (1993), who studied line drawings of objects. Because of the apparent dependence on the visible parts, Biederman and Gerhardstein had interpreted their findings as evidence in favor of part-based structural descriptions. Interestingly, Hayward found the same result when the second image was a silhouette. This finding can be interpreted in one of two ways; either the information in the silhouettes was sufficient to recover the object's parts, or the similarity in the shapes of the bounding contours in the two intervals was used directly to perform the task. The later possibility would not implicate the use of a part-based representation of shape.

The possibility that the bounding contour is used directly as a visual code is suggested by the success of a number of artificial recognition systems. Richards and his colleagues (Richards, Dawson, & Whittington, 1986; Richards, Koenderink, & Hoffman,

1987) developed a scheme in which the bounding contour of an object is described as an ordered list of curvature primitives termed codons. Codons capture qualitative properties of the curvature of the bounding contour, which makes the representation invariant to image scaling.

Ullman (1996) discusses a similar procedure used by Yolles to classify images of objects—including novel exemplars of known object classes—using only information contained in the bounding contour. In this scheme, the bounding contour is decomposed into a collection of small elements, each of which has a 2-D position and orientation. The "distance" between any two such elements depends on the difference between both their positions and their orientations. Using this distance measure, it is possible to find the stored image that matches best any input image, thus making classification of even novel images possible. Alternative direct coding methods have also been developed that utilize fourier descriptors to describe image contours (e.g., Zahn & Roskies, 1972).

Regardless of the actual code used, however, if the shape of the bounding contour itself is used as a visual code, then silhouette images should always be recognized as well as shaded images, so that view canonicality effects should be similar for silhouettes and shaded images.

There are reasons, however, to doubt that the shape of the bounding contour is itself used as a visual code. For example, although three-quarter views of objects are readily recognized, it is difficult to imagine or to draw the shape of the bounding contour of objects in three-quarter views. Consider how difficult it would be to draw a car or a cow seen in a three-quarter view. In contrast, it is much easier to imagine or to draw the bounding contour of a car or a cow seen in profile. This is not to say that one cannot imagine what an object looks like in a three-quarter view. In fact, Blanz et al. (1996) found that observers, when asked to imagine an object, most often imagined the object in a three-quarter view. Thus, there may be features present in the three-quarter view that makes it easy to recognize and imagine, but the bounding contour might not be one of these features. Subjectively, the difference between the three-quarter view and the profile view is that the object's parts can be better inferred from the bounding contour in the profile view. It might be that the bounding contour is sufficient for recognition only when it allows for the recovery of other features, for example, the object's primary axis of elongation or the object's parts.

Hoffman and Richards (1984) investigated formally whether an object's bounding contour could be used to recover a part-based description of the object. They argued on the basis of geometrical principles that the locations of some of an object's part boundaries could be inferred from the locations of discontinuities or extrema in the curvature of its bounding contour. As the salience of these features is likely to vary with the viewing direction, one might still expect to find view effects in recognition if an object's parts were recovered in this way. It is not clear, however, whether one would expect to find differences in view effects for shaded and silhouette images. If part boundaries were also inferred in shaded images on the basis of shading gradients internal to the object's bounding contour, then view effects might be very different for shaded and silhouette images.

In the present experiments, we will examine directly whether canonical view effects are the same for shaded and silhouette images of objects. In Experiment 1, observers will name shaded and silhouette versions of four-legged animals and artifacts (i.e., non-animals) seen in three different views, *three-quarter*, *profile*, and *back* views. Three-quarter and profile views have led to fast naming times in previous studies, whereas naming times for views from the back have been found to be reliably slower. We will examine the results for artifacts and animals separately, because the animals, being more similarly shaped, could show canonical view effects that are different from

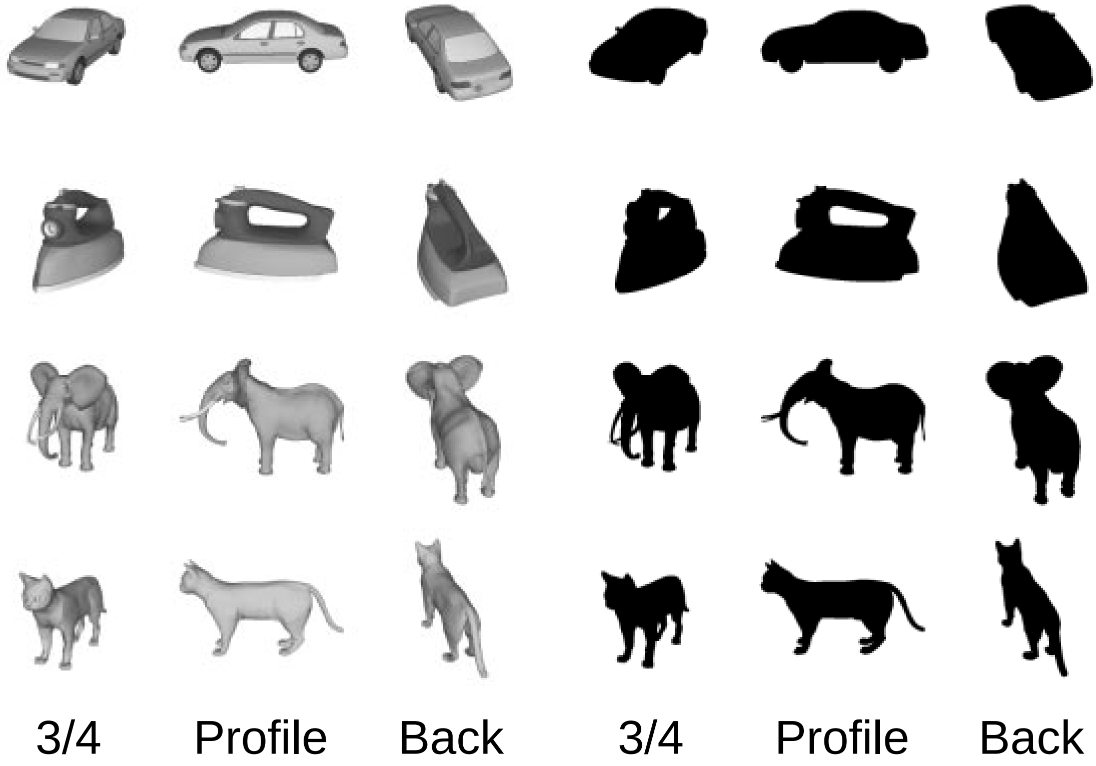

Figure 1: Shaded and silhouette images of four of the objects studied in Experiment 1. Three-quarter, Profile, and Back views of each object are shown.

those exhibited by the artifacts. In Experiment 2 we will present ideal-observer analyses of the images used in Experiment 1 to examine whether there are inherent differences in the complexity of the stimuli used in the various conditions.

2 Experiment 1: Object Naming

2.1 Method

Observers. The observers were 12 student volunteers from Eberhart-Karls University in Tübingen, Germany. Each received a payment of 6 DM. All were native speakers of German and reported having normal or corrected-to-normal visual acuity.

Stimuli. The stimuli were 256 pixel by 256 pixel images of 18 common objects, 9 artifacts and 9 four-legged animals. The objects were three-dimensional computer graphics models acquired from Viewpoint Data Labs.[1] Each object model was imaged from three different viewpoints using custom Silicon Graphics Inventor software that simulated a virtual perspective camera located 50 cm from the object. Each object was scaled to fit within a sphere of radius 7 cm and was oriented so that in the "home" view it faced the camera. For elongated objects, the primary axis of elongation was aligned with the direction of the camera in this view. The *Three-quarter* view of each object was generated by elevating the camera 15.2° and subsequently rotating it 26.7° about the vertical axis. The *Profile* view was generated by elevating the camera 15.9° and rotating it 90.0° about the vertical axis. Finally, the *Back*

[1] Some of the models are available from Viewpoint's World Wide Web page at http://www.viewpoint.com. All of the objects can be purchased from Viewpoint. Images of some of the objects can be downloaded from the MPI home page at http://www.mpik-tueb.mpg.de.

view was generated by elevating the camera 30.0° and rotating it 159.1° about the vertical axis. These particular viewing angles were chosen to match some of the viewing angles used in a separate series of experiments (Liter & Schölkopf, 1998).

Two versions of each view of each object were generated by changing the surface material properties of the object model prior to rendering. *Shaded* versions of each view were created by coloring the surfaces of the object grey. *Silhouette* versions were created by coloring the surfaces black. In both cases the specular component of the surface material was set to zero, so that none of the images contained distinctive highlights. The objects were illuminated by an omnidirectional ambient light source and a directional light source located 45° above and to the right of the camera. The rendering model used Gouraud shading. Because the black-surfaced objects did not reflect any light, they appeared as black-on-white silhouettes even when lit. Examples of the Shaded and Silhouette versions of some of the objects are shown in Figure 1.

Apparatus. The experiment was conducted using a Power Macintosh 9500/132 running PsyScope version 1.1 experimental design software (Cohen, MacWhinney, Flatt, & Provost, 1993). Subjects' naming responses were triggered by a Yoga EM-240 electret condenser microphone and registered to the computer via a CMU button box. A cassette tape recorder (Sony CFS-B11) was used to record the observers' verbal responses.

Procedure. The observers participated individually in 20-min sessions. Each observer read printed instructions explaining that the task was to name pictures of objects and animals that would appear on the computer monitor as quickly and as accurately as possible using the first name that came to mind. Each trial began with the presentation of a small ready prompt. The observer initiated the trial by pressing a button on the button box. The picture appeared 750 ms after the button press and remained visible until the microphone registered a response. The ready prompt for the next trial appeared after an additional 500 ms delay. Before beginning the experimental trials, the observers completed 20 practice trials. The practice objects were all different from the experimental objects.

Design. There were four independent variables in this experiment, a) Object Type (Artifact or Animal), b) Viewpoint (Three-quarter, Profile, or Back), c) Type of Rendering (Shaded or Silhouette), and d) Block (1-6). Each observer named all 18 objects in each block. Within each block, three objects were seen shaded in three-quarter views, three were seen as silhouettes in three-quarter views, and so on. The six sets of three objects were maintained across observers and blocks, which made it possible to counterbalance the assignment of objects to the various conditions across blocks. Each set of three objects contained both artifacts and animals. For any given observer the viewpoint and type of rendering for each set of three objects was different in every block. Across observers, each set of three objects was seen twice in each block in each of the six viewpoint and rendering conditions.

2.2 Results

Analyses of response times were carried out only on correct responses. The experimenter scored each naming response as correct or incorrect by listening to the tape recording that accompanied each observer's experimental session. Questionable responses were referred to one or more judges. Trials in which the observer named the object correctly at a level other than the entry level were scored as correct. By listening to the tape recordings, it was also possible to identify trials in which the observer made false starts and trials in which there were equipment failures. These trials were scored as errors. In all, 56 of the total 1296 trials (4.3%) were scored as errors. The distribution of errors across Object Type, Viewpoint, and Type of Rendering will be presented below. After removing error trials, each observer's mean response time was computed. Response times that differed

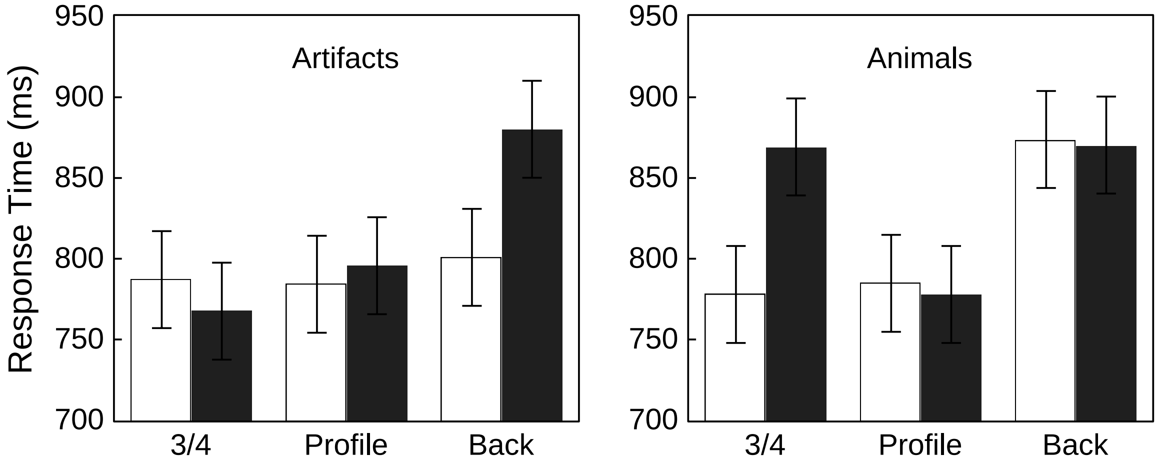

Figure 2: Mean naming response times in block 1 of Experiment 1 for the shaded (open bars) and silhouette (dark bars) versions of each view. Error bars are within-subjects standard errors (see Loftus & Masson, 1994).

from an observer's mean by more than two standard deviations were then removed. This amounted to 66 of the 1240 correct responses (5.3%).

Because of the potential influence of repetition priming effects (both visual and nonvisual) in later blocks of the experiment, the naming response times for each block were analyzed separately. As each object was seen for the first time in block 1, performance in this block allows us to examine whether viewpoint and type of rendering influence the initial recognition of each object and whether these effects are different for artifacts and animals.

The mean block 1 naming times for the various experimental conditions are shown in Figure 2. Naming times were relatively fast for all three views of the shaded artifacts. In contrast, there were differences in naming times across views when the artifacts were seen as silhouettes. Silhouette artifacts were named as quickly as shaded artifacts in three-quarter and profile views, but they were named more slowly in back views. The pattern of results for animals was somewhat different. Unlike the artifacts, response times for shaded animals did vary with viewpoint. Three-quarter and profile views were named relatively quickly, whereas back views were named more slowly. The pattern of results was again different for silhouettes. Animals seen as silhouettes were named quickly only in profile views.

These results indicate that performance is not always equivalent for shaded and silhouette images of objects. Artifacts were named more slowly when they were seen as silhouettes in back views, and animals were named more slowly when they were seen as silhouettes in three-quarter views.

Performance in block 2 was surprisingly similar to that in block 1. One might have expected that because of visual and nonvisual priming effects the differences in naming times across viewing conditions would have been substantially diminished or even eliminated in block 2. In contrast, it appears that the absolute effects produced by the various viewing conditions had a greater effect on naming times in block 2 than did any priming effects. If we consider how naming times changed from block 1 to block 2 for particular objects, it is clear that the differences in naming times reflect differences in the viewing conditions rather than general priming effects. For example, silhouette artifacts seen in profile in block 1 were named 34.6 ms faster in block 2 when they were seen in three-quarter views. Given the same changes in viewing conditions, silhouette versions of animals were named 244.4 ms *slower* in block 2. These differences reflect the finding in block 1 that both profile and three-quarter views of artifacts were easily recognized, whereas only profile views of animals were easily recognized. Such contrasts were also found for other transfer conditions. For example, shaded artifacts seen from the back in block 1 were named 8.8 ms faster in block 2 when they were seen in profile, whereas shaded animals were named 142.6 ms faster given the same changes in viewing conditions. Again, these differences reflect the differences in absolute naming times observed for the various conditions in block 1.

The error rates in this experiment were quite low, but they generally conformed to the pattern followed by the response times in the early blocks of the experiment. The number of errors (summed over all observers and all six blocks of trials) for three-quarter, profile, and back views, respectively, were 2, 1, and 1 for shaded artifacts, 9, 2, and 11 for silhouette artifacts, 1, 2, and 9 for shaded animals, and 4, 4, and 10 for silhouette animals.

2.3 Discussion

The results for shaded objects replicate those found in previous naming studies carried out with these objects (Liter & Schölkopf, 1998), and they are in general agreement with other studies of canonical view effects (Blanz et al., 1996; Palmer et al., 1981). Three-quarter and profile views were recognized relatively quickly, whereas views from the back were sometimes recognized more slowly. The results for the silhouette images, however, did not in all cases follow those for the shaded images. In particular, three-quarter views of animals were not recognized quickly when they were shown as silhouettes, although they were recognized relatively quickly when shaded. This was also true for the back views of the artifacts. These views were recognized as quickly as three-quarter and profile views when they were shaded, but they were not recognized quickly in silhouette.

These results provide evidence *against* the hypothesis that the bounding contour is used directly as a visual code. Rather, they suggest that the bounding contour supports efficient recognition only when it is possible to recover other features from the bounding contour that are themselves elements of the actual long-term visual representation. What these features might be cannot be determined from the present experiment. Nevertheless, a few observations might shed some light on this question. The views for which recognition was efficient with silhouette images were generally those views that clearly showed the parts of the object or certain of the object's distinctive features. For example, in the profile view of each animal it was generally possible to recover from the bounding contour the aspect ratio of the animal's body and the lengths of its legs relative to the overall size of its body. Furthermore, it was often the case that certain distinctive features were visible in the profile view that were less visible in other views. For example, the horn of the rhino was clearly visible in the profile view as was the curly tail of the pig and the hump on the camel's back.

At first sight, these results might seem to be at odds with those of Hayward (1995), who found that performance with silhouette images was almost always as good as it was with shaded images. The discrepancy could stem from differences between the tasks used by Hayward and the task used in the present experiment. All of Hayward's tasks involved some sort of stimulus repetition. In Hayward's matching experiments, observers saw a shaded view of an object in a canonical view and then decided whether a subsequently presented view, which was either shaded or silhouette, showed the same object. Similarly, in Hayward's naming experiments, observers named a sequence of objects seen in shaded canonical views and then named them again in a second sequence in either shaded or silhouette views. To perform well in these experiments, it might only have been necessary to make contact with the representation that was established the first time an object was seen. It might be much easier to use the information contained in a silhouette for this purpose compared to how it must be used to support recognition of an object never before encountered.

The absence of a substantial priming effect between blocks 1 and 2 in the present experiment is somewhat inconsistent with this explanation. It should be noted, however, that by the sixth block of the experiment, most of the effects of viewpoint and type of rendering had vanished. Thus, there probably were subtle priming effects from block to block, but they were masked in early blocks by the larger effects that resulted from differences in the viewing conditions.

One might also attempt to explain the failure to observe priming in block 2 by arguing that priming depends on the quality of the view that is seen in block 1. A shaded, three-quarter view might provide a strong visual representation, capable of supporting priming, whereas a silhouette view of the back of an object might not provide a strong enough representation. Each observer in the present experiment saw only three objects in a shaded three-quarter view in block 1. Thus, if priming depends on the goodness of the view seen, then there would have been little chance of priming in block 2. After several blocks, however, the visual representation of each object might have improved enough to support priming and minimize the effects of the various viewing conditions. Although such an argument has not been made in the priming literature, it might be useful to consider. For example, Hayward (1995) did observe priming in his naming experiments, but in all cases observers viewed shaded objects in canonical views in block 1.

3 Experiment 2: View Complexity Measurements

To interpret the results of Experiment 1 as reflecting processing limitations of the human visual system, it is first necessary to establish that the differences observed in Experiment 1 cannot be accounted for by differences in the complexity of the stimuli used in the various conditions. It is possible, for example, that the silhouette images were intrinsically less distinctive than the shaded images. Likewise, images of the backs of the objects might have been less distinctive than the three-quarter or profile views.

Tjan and Legge (in press) developed an objective procedure based on ideal observers to measure the complexity of different object recognition tasks. The procedure determines the number of 2-D images of each object (termed the view complexity or VX) in a specified set of objects that must be stored so that any object in the set can be identified from any 3-D viewpoint. An important property of the procedure is that it does not assume the use of any particular form of visual representation (i.e., 3-D models or structural descriptions). The 2-D images of the objects themselves serve as the representation. Because of this, it is possible to attribute differences in view complexity for different sets of stimuli to intrinsic differences in the stimuli rather than to the success or failure of some unknown process that attempts to derive a representation from the stimuli. For example, Tjan and Legge (in press) found that recognition tasks involving geons (such as those used by Biederman & Gerhardstein, 1993) are less complex than those involving bent wire objects (such as those used by Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992). This suggests that differences in viewpoint dependence observed for these two classes of objects might not reflect any particular processing strategy of the human visual system but rather the intrinsic differences in the complexity of the stimuli.

The view complexity of a set of objects is assessed in a numerical simulation by gradually increasing the number of views of each object that must be identified and measuring for each number of views the threshold signal-to-noise ratio (SNR) that yields 90% correct identification. The views that must be identified are termed "task" views, as they are the only views that must be recognized. SNR is manipulated by adding Gaussian noise to each pixel of the display. Tjan and Legge (in press) showed that the threshold SNR increases linearly with the log of the number of views and then reaches a plateau. The number of views per object at which the threshold SNR reaches a plateau, as determined by a bi-linear fit, is defined to be the view complexity of the object set.

The increase in threshold below the view-complexity boundary indicates that the task becomes more difficult as more views of the objects must be identified. This means that the additional views are sufficiently distinct from the views already stored that they provide further information about the identities of the different objects. Beyond the view-complexity boundary, additional views do not provide more information, so that the objects can be thought of as being completely represented at the view-complexity boundary.

The view complexity analyses performed on the stimuli studied in Experiment 1 were similar to the analysis described in Tjan and Legge (in press), but they were modified to allow an assessment of the view complexity of specified regions of the viewing sphere. On the basis of the results of Experiment 1, we were particularly interested in whether view complexity was greater for views in the neighborhood of "back" views than it was for views in the neighborhood of "three-quarter" views. Likewise, we were interested in comparing the complexity of shaded and silhouette images.

3.1 Method

Ideal observer analyses were performed to assess view complexity for eight of the stimulus conditions studied in Experiment 1. Profile views were not examined, because human performance was similar in all conditions for these stimuli. The task viewpoints, which were the same for all analyses, were selected by uniformly and randomly sampling 24576 viewpoints from the upper half of the viewing sphere. The number of task views was 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, or 24576. The first 8 viewpoints in the list of 24576 task views comprised the 8 view-per-object condition. The first 16 viewpoints comprised the 16 view-per-object condition, and so on. For each number-of-views condition, one image of each object was rendered from each task viewpoint.

Tjan and Legge (in press) demonstrated that it is not necessary to test recognition at all of the task viewpoints to achieve a stable measure of view complexity. Rather, it suffices to conduct the view complexity analysis several times with independent samples of the task viewpoints. The results of these analyses can then be averaged. In the analyses performed here, one test viewpoint was selected from the set of task viewpoints, and this viewpoint was used to render an image of each of the nine objects. These nine images served as test images throughout the analysis. This procedure was repeated eight times, each time with a new test viewpoint (and thus 9 new test images).

An important difference between Experiment 1 and the ideal observer developed by Tjan and Legge (in press) is that the observers in Experiment 1 were tested on only certain views of the objects, whereas Tjan and Legge's ideal observer was tested on arbitrary views. Although the "observers" in both cases did not know in advance which views would be tested, restricting the test views in Experiment 1 might have led to differences in performance because the stimuli were more or less distinctive in the different views. To better equate the human and ideal observer tasks, we modified the ideal observer procedure used by Tjan and Legge. Rather than select the test viewpoint randomly from the set of task viewpoints, we selected the test viewpoint with the constraint that it fell within a cone of radius 15 deg centered at either the three-quarter viewpoint or the back viewpoint (depending on the analysis).

Because it was unlikely that any of the first 8 task viewpoints fell within the specified cone, it was necessary to continue sampling from the ordered list of task viewpoints until a viewpoint was found that did fall within the cone. This viewpoint was then added to the task view sets that did not already contain that viewpoint. For example, if no task

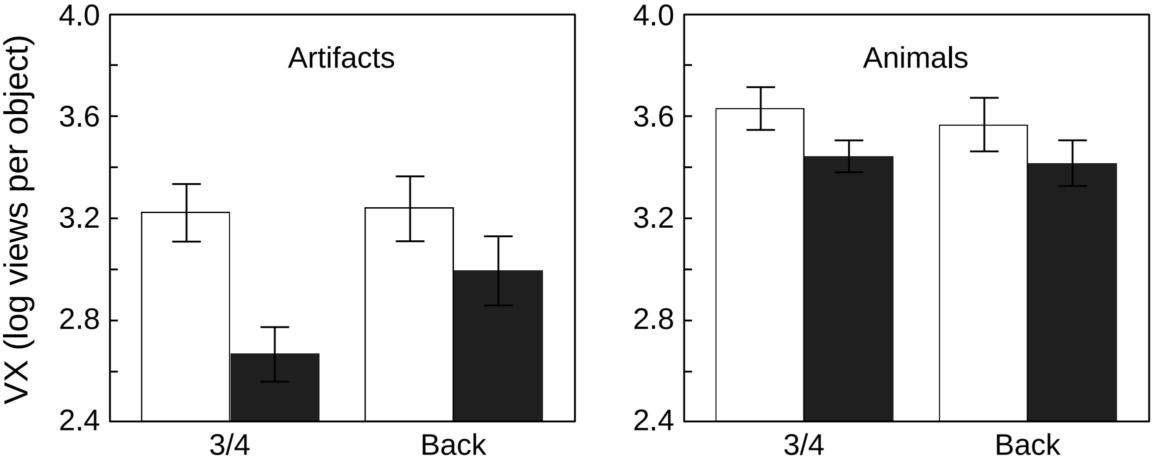

Figure 3: View complexity expressed in log number of views per object for the shaded (open bars) and silhouette (dark bars) versions of three-quarter and back views. Between-subjects standard error bars were computed separately for each condition by running the simulation with 8 independent samples of test views.

viewpoint lying within the specified cone was found in the first 16 task viewpoints, but one was found in the first 32 task viewpoints, that viewpoint was added to the 8 view-per-object and 16 view-per-object sets. The benefit of using this procedure to select test viewpoints within specified regions of the viewing sphere was that it did not make the overall distribution of task views less uniform over the viewing sphere.

3.2 Results and Discussion

The results of the view complexity analyses are shown in Figure 3. Contrary to what one might have expected on the basis of the Experiment 1 results, in all cases view complexity was found to be greater for the shaded images than for the silhouette images. Although this result might seem unintuitive, it is easily explained by the fact that the view complexity algorithm attempts to represent all of the information that is present in the images. There are simply more details, and thus there is more to represent, in the shaded images. The finding that human subjects' performance is opposite to that expected from view complexity indicates that performance differences are not due to intrinsic differences in the distinctiveness of the stimuli in different conditions. Rather, differences in performance reflect actual differences in the ability of the human visual system to construct adequate representations from the stimuli.

With the possible exception of the silhouette versions of the artifacts, the view complexity analyses failed to show reliable differences between the three-quarter and back views. The back views of these objects are as distinctive as the three-quarter views. The difference in recognition performance between these conditions in Experiment 1 indicates that the human visual system is better able to derive a useful representation from three-quarter views.

The view complexity analyses also indicated that the animals were more complex than the artifacts. Although this was not apparent in Experiment 1, we have found in other experiments with these objects that the animals do take longer to name on average than do the artifacts (Liter & Schölkopf, 1998). Why this was not observed in Experiment 1 is unclear.

4 General Discussion

The purpose of this study was to examine whether the equivalence observed in previous experiments for shaded and silhouette images of objects would also be observed in object naming. Hayward (1995) found that silhouettes served as well as shaded images in object comparison and priming tasks. Experiment 1 presented here, however, demonstrated that silhouette images do not serve as well as shaded images in object naming in some important cases. One implication of this finding is that the shape of an object's bounding contour does not serve as a direct visual code for the object, at least not exclusively. Rather, it appears that the visual system relies on some other representation that can be derived easily from the bounding contour in some views but not in others.

In some of the views in which the bounding contour does not suffice for efficient recognition, shading provides additional information that does allow for efficient recognition. Although it cannot be determined from Experiment 1 precisely which of the features provided by shading are being used to enhance recognition, there are many interesting possibilities. For example, shading indicates that the object depicted is not flat. Furthermore, as discussed by Newell and Findlay (in press), shading can also provide information about the orientation of the object, which could be useful in assigning an internal, object-based reference frame to the object.

Shading also provides some information about surface structure, such as the relative depths of surface patches, and it provides information about distinctive features and their relative positions. For example, shading might enhance recognition of the animals in three-quarter views because it makes the animals' facial features and the spatial relationships among the features apparent. Such features might themselves be elements of the long-term visual representation, as would be the case in models of long-term visual memory that posit storage of rich, view-based descriptions of objects (Bülthoff et al., 1995).

Shading might also be used to assist in the derivation of a part-based representation of the object. In certain views such as the profile view, it might be relatively easy to derive a part-based description from the bounding contour alone. As pointed out by Hoffman and Richards (1984), it is often possible to correctly infer the locations of part boundaries by locating discontinuities or extrema in the curvature of the bounding contour. In other views, it might not be possible to derive a part description from the bounding contour that is sufficient to identify the object. In these views, shading gradients or discontinuities might be used to provide a more accurate or complete part-based description. Shading could indicate that a relatively short protrusion on an object's boundary is actually a much longer part that extends into the interior of the bounding contour. Shading could also make other parts visible that do not protrude at all from the boundary of the object. In Figure 1, for example, the rear windshield of the car is not visible in the silhouette image of the back view of the car. It is, however, quite obvious in the shaded image.

An alternative hypothesis concerning the role of the bounding contour in recognition is that it serves as a first index into visual memory, limiting the space of alternatives items that must be considered before recognition can take place. In some cases, this first index might be sufficient to perform the required task. For example, if observers in Experiment 1 were only required to determine whether the objects presented were artifacts or animals, the bounding contour alone likely would have been sufficient in all views. For profile views of animals, the first index provided by the bounding contour might also be sufficient to identify the particular animal that is depicted. To identify particular animals in three-quarter and back views, however, additional processing is required beyond the first index. The hypothesis that processing is hierarchical and task-dependent was discussed by Rock and DiVita (1987), and some researchers have recently developed these ideas further, suggesting that recognition proceeds from global to local aspects of shape (e.g., Sanocki, 1993).

In summary, we have found that shading can provide important information for efficient object recognition. A question for future research will be to determine more precisely which of the many features provided by shading are actually used to enhance recognition. Shading could be used to derive a better representation of the object's parts, or it might be used to extract localized features that are themselves elements of the long term visual representation.

References

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

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.

Blanz, V., Tarr, M. J., Bülthoff, H. H., & Vetter, T. (1996). *What Object Attributes Determine Canonical Views?* (Tech. Rep. No. 42). Tübingen, Germany: Max-Planck-Institute for Biological Cybernetics.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, *89*, 60-64.

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

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). Psyscope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments & Computers*, *25*, 257-271.

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.

Hayward, W. G. (1995). Effects of outline shape in recognition. Manuscript submitted for publication.

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition*, *18*, 65-96.

Liter, J. C., & Bülthoff, H. H. (1997). *View Canonicality Affects Naming but not Name Verification of Common Objects* (Tech. Rep. No. 51). Tübingen, Germany: Max-Planck-Institute for Biological Cybernetics.

Liter, J. C., & Schölkopf, B. (1998). *Psychophysical and Computational Experiments on the MPI Object Databases* (Tech. Rep. in preparation). Tübingen, Germany: Max-Planck-Institute for Biological Cybernetics.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476-490.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional shapes. *Proceedings of the Royal Society of London, B*, *200*, 269-294.

Newell, F. N., & Findlay, J. M. (in press). The effect of depth rotation on object identification. *Perception.*

Palmer, S. E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance ix* (p. 135-151). Hillsdale, NJ: Lawrence Erlbaum Associates.

Perrett, D. I., Smith, P. A., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1984). Neurones responsive to faces in the temporal cortex: studies of functional organization, sensitivity to identity and relation to perception. *Human Neurobiology*, *3*, 197-208.

Richards, W., Dawson, B., & Whittington, D. (1986). Encoding contour shape by curvature extrema. *Journal of the Optical Society of America A*, *3*, 1483-1491.

Richards, W. A., Koenderink, J. J., & Hoffman, D. D. (1987). Inferring three-dimensional shapes from two-dimensional silhouettes. *Journal of the Optical Society of America A*, *4*, 1168-1175.

Rock, I., & DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology*, *19*, 280-293.

Sanocki, T. (1993). Time course of object recognition: Evidence for a global-to-local contingency. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 878-898.

Tjan, B. S., & Legge, G. E. (in press). The viewpoint complexity of an object-recognition task. *Vision Research.*

Ullman, S. (1996). *High level vision.* Cambridge, MA: MIT Press.

Zahn, C. T., & Roskies, R. Z. (1972). Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, *C-21*, 269-281.