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

Technical Report No. 51

# View Canonicality Affects Naming But Not Name Verification Of Common Objects

Jeffrey C. Liter[1] & Heinrich H. Bülthoff[2]

June 1997

Presented at the Annual Meeting of the Association for Research in Vision and Ophthalmology, Ft. Lauderdale, FL, May 11-16, 1997. This research was supported by a grant from the Human Frontier Science Program to J. Liter. The authors would like to thank V. Blanz for assistance in selecting and producing the stimuli used in these experiments.

[1] AG Bülthoff, E–mail: jliter@mpik-tueb.mpg.de
[2] AG Bülthoff, E–mail: hhb@mpik-tueb.mpg.de

This document is available as /pub/mpi-memos/TR-051.ps via anonymous ftp from ftp.mpik-tueb.mpg.de or from the World Wide Web, http://www.mpik-tueb.mpg.de/bu.html.

# View Canonicality Affects Naming But Not Name Verification Of Common Objects

*Jeffrey C. Liter & Heinrich H. Bülthoff*

**Abstract.** Previous studies suggest that there are "canonical" viewpoints from which objects are identified most readily. Object naming has been the primary objective measure used to assess view canonicality, but this task has not proven adequate in distinguishing the many explanations of this phenomenon that have been offered. We examine object naming (Experiment 1a) and name verification (Experiments 1b and 2) to better understand the origin and nature of canonical view effects in recognition. In the name verification experiments, observers read an object name and then viewed an image of an object and decided as quickly as possible whether the image matched the name. The stimuli were images of 3D computer models of seven common objects. Each object was rendered from one canonical viewpoint (determined in a separate experiment by Blanz, Tarr, Bülthoff, & Vetter, 1996) and two noncanonical viewpoints. Observers named the objects faster in canonical views, but performance was not affected by viewpoint in either name verification experiment, even on the first presentation of each view. We interpret these results in terms of a view-based similarity model. Naming is slow for noncanonical views because they are similar to stored views of more than one object, leading to response competition. The name verification task reduces the space of relevant views in long-term memory that must be compared to an input view, which for the views studied here minimized the likelihood of confusions and eliminated differences in response times for different views.

## 1 Introduction

Although it might not be apparent in our everyday subjective experience, the viewpoint from which an object is seen can sometimes influence how easily it is recognized. Palmer, Rosch, and Chase (1981) studied this phenomenon extensively in a series of experiments in which they measured among other things the subjective goodness of different views of objects and the time needed to name objects seen from different viewpoints. Views that were rated as subjectively good views were named more quickly than views rated not as good. Palmer et al. (1981) termed these good views "canonical" views. It is perhaps not entirely surprising that different observers have similar criteria for determining what constitutes a good or a bad view of an object, but that the perceptual processing of these views is different is intriguing. Why should some views of objects be recognized more quickly or processed more efficiently than other views?

Explanations of canonical view effects differ according to the way in which objects are believed to be represented in long-term visual memory. For this reason, a better understanding of when and why canonical view effects occur is important to gain insight into our visual representation of objects. We will consider two classes of models describing the visual representation of objects, 1) those in which each object is represented by multiple viewpoint-specific descriptions (Biederman, 1987; Bülthoff & Edelman, 1992; Bülthoff, Edelman, & Tarr, 1995; Tarr, 1995; Perrett, Smith, Potter, Mistlin, Head, Milner, & Jeeves, 1984; Ullman, 1989), and 2) those in which each object is represented by a single three-dimensional, object-centered structural description (Lowe, 1987; Marr, 1982; Marr &

Nishihara, 1978).

An important property of multiple-view models of long-term visual memory is that only a relatively small number of views need be stored for any given object.[1] Storing all views of an object would be impossible, as there are uncountable 3D viewpoints from which any object could be seen. Differences in speed or accuracy of recognition of different views of an object might then result for one of several reasons. It is possible, for example, that different stored views are weighted differently, such that especially important views, or views that are encountered often, are weighted most heavily. Repeated processing of the same or similar views of an object view could lead to more efficient processing of that view, perhaps because a greater number of neurons are recruited to support recognition of that view (e.g., Perrett, Oram, & Wachsmuth, 1996). Explanations of view canonicality in terms of familiarity or frequency of exposure are consistent with Palmer et al.'s (1981) finding that the visibility of an object's front surface accounted for over 50% of the variance in the goodness ratings their observers gave to different views of the object. This finding supports familiarity-based explanations in as much as the front surface of an object is likely to be the surface interacted with (and thus experienced) most.

Differences in the speed or accuracy of recognition of different views of an object could also arise because some views are not represented and can be recognized only through the use of a time consuming or error prone transformation mechanism. For example, theorists have proposed that views that are not represented might be recognized by interpolating among views that are represented (e.g., Poggio & Edelman, 1990) or by transforming either the input view, the stored views, or both (e.g., by mental rotation, Shepard & Cooper, 1982, Tarr, 1995, or by alignment, Ullman, 1989, 1996). It is reasonable to expect that canonical views would correspond to those views that do not need to be transformed or need to be transformed less to be recognized.

Theorists who have developed models of long-term visual memory based on object-centered structural descriptions have argued that canonical view effects occur because an object's structural description cannot be generated as easily from all views of the object. According to these models, objects are recognized by constructing a three-dimensional structural description of a viewed object and then matching the description to descriptions stored in memory (Lowe, 1985, 1987; Marr & Nishihara, 1978; Marr, 1982). Of primary importance in these models is that the descriptions are orientation free. This is because the descriptions are based on a reference frame that is internal to the object. In Marr's model, for example, the parts of an object are coded relative to its main axis of elongation. Because the stored models are orientation free, the ease with which a description derived from an image can be matched to a stored description should not depend on the viewpoint from which the object was seen.

How then can one account for canonical view effects in these models? Consider what would happen if the "backbone" of the description, for example, the object's main axis of elongation, were difficult or impossible to recover when the object was seen from certain viewpoints. This would disrupt construction of a structural description, making recognition of the object in these views difficult or impossible. For these structural-description models, then, the canonicality of a given view should be a function of the ease with which a three-dimensional structural description can be generated from the view and not the familiarity of the view.

The results of several empirical studies of object recognition are consistent with this axis-foreshortening interpretation of canonical view effects. Lawson (1994) found that ob-

[1] We will use the term "view" to refer to both the image of an object and its internal description. However, we are not suggesting that internal descriptions are image-like templates or "pictures in the head." Tarr (1995) presents a nice discussion of why view-based descriptions are not templates.

servers took longer to name objects if they were shown from a viewpoint in which the main axis of the object was foreshortened (i.e., parallel to the line of sight). Newell and Findlay (1996) found a similar difficulty with foreshortened views in a name verification task in which the name of the object was displayed together with an image of the object. Humphrey and Jolicoeur (1993) found that the time needed to name foreshortened views of objects could be lessened if pictorial depth cues were added to the display. This suggests that at least some of the difficulty associated with identifying objects seen in foreshortened views is that the depth perceived in these views is incorrect.

One could interpret these results as evidence for *non*canonical view effects in recognition rather than canonical view effects. Nevertheless, the degree of axis foreshortening is unlikely to be the only determiner of view canonicality. Palmer et al. (1981) found a monotonic increase in naming times for a wide range of views, not just a difference for nearly or fully foreshortened views. To be sure that any view effects that might be observed in the present study are not simply due do difficulties in recognizing foreshortened views, we will avoid such views.

One weakness of much of the previous research on view canonicality is the limited use of objective measures. Researchers have used subjective measures such as goodness ratings (Palmer et al., 1981) and exploration times (Perrett & Harries, 1988; Harries, Perrett, & Lavender, 1991; Perrett, Harries, & Looker, 1992), but the primary objective measure has been object naming. This stands in stark contrast to research on other viewpoint effects in object recognition, for example, how a change of viewpoint between study and test episodes affects recognition performance. Researchers studying the effects of change of viewpoint have used many different objective tasks, including object naming (Biederman & Gerhardstein, 1993), old-new recognition (Edelman & Bülthoff, 1992; Liter, 1996), and picture-picture comparison (Bartram, 1976; Ellis & Allport, 1986).[2] The complexity of the data that has emerged from this body of research suggests that it would be unwise to assume that because canonical view effects occur in object naming they will occur in the same way for other tasks.

In the present study we examined canonical view effects in a wider range of tasks by studying object naming and name verification with the same set of stimuli. In the name verification task, observers read an object name then decided whether a subsequently presented object matched the name. The objects and views used in the present study were based on a study by Blanz, Vetter, Bülthoff, and Tarr (1995) in which observers actively rotated 3D computer models of objects using a "Spaceball," a mouse-like input device with three rather than two degrees of freedom. In one of their experiments, observers selected views of the objects that they believed would be best for displaying the objects in a brochure. In another experiment, observers first generated a mental image of an object then produced that view by rotating the computer model. The views selected in the brochure task were used in the present study as the canonical views. Noncanonical views were chosen by rotating the objects about the vertical axis and viewing them from higher elevations. Accidental views in which the main axis of the object was appreciably foreshortened were avoided.

### 1.1 Experiment 1a

The primary purpose of this experiment was to establish that the canonical views derived from the Spaceball experiments of Blanz et al. (1995) were in fact canonical in terms of naming response times. Observers named each of seven objects in three views. The same observers participated in this experiment and in Experiment 1b, in which a name verification procedure was used.

[2] Surprisingly, with the exception of studies examining foreshortened views, little attention has been given to the goodness of the views used in tasks examining the effects of change of view.

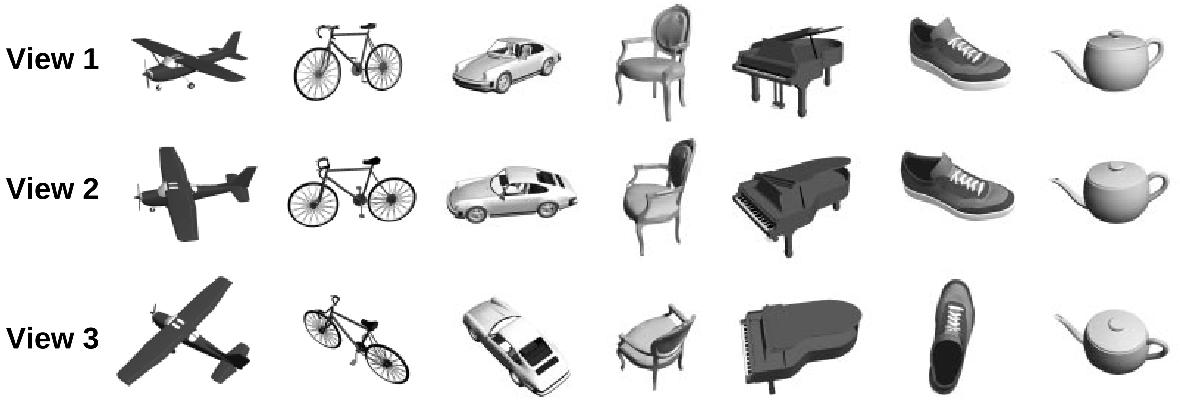

Figure 1: The images used in all three experiments. View 1 corresponds to the “canonical” view found by Blanz et al. (1995). Views 2 and 3 depict each object from less canonical viewpoints.

### 1.1.1 Method

**Observers**. Seven student volunteers from Eberhard-Karls University in Tübingen, Germany participated in the experiment for a payment of 10 DM. All were native German speakers and reported having normal or corrected to normal visual acuity.

**Stimuli.** The stimuli were 21 images generated from three-dimensional computer models[3] of seven familiar objects (airplane, bicycle, car, chair, piano, shoe, & teapot).[4] Three 512 by 512 pixel images of each object model were generated using custom SGI Inventor software that simulated a virtual camera located 50 cm from the object. The objects were scaled to fit within a sphere of radius 7 cm. The direction of view was different for each image, as shown in Figure 1. The surfaces of the objects were colored with various shades of gray. The objects were illuminated by an omnidirectional ambient light source and a directional light source located 45 degrees above and to the right of the camera. The rendering model used Gouraud shading. The specular component of the surface material was set to zero so that the images did not contain distinctive highlights.

Each object was oriented so that in the “zero” view it faced the camera and its main axis of elongation pointed toward the camera. The teapot’s spout pointed toward the camera in this view. The “canonical” view (View 1) of each object was determined in an experiment reported by Blanz et al. (1995) in which observers rotated the computer models in real time using a “Spaceball” and selected the view they thought would be best for displaying the object in a brochure. View 1, which was slightly different for each object, was roughly a “three-quarter” view seen from approximately 15 degrees of elevation and with the main axis of the object rotated approximately 45° about the vertical axis. View 3 depicted the object from a higher elevation and more from the back. The viewing direction used to generate View 2 was exactly half way between the viewing directions used to generate Views 1 and 3 on a virtual viewing sphere surrounding the object. The complete set of images used in these experiments is shown in Figure 1.

**Apparatus.** The experiment was conducted using an SGI Indigo 2 workstation equipped with a 24 bit High Impact graphics card and a 35.2 cm wide (1280 pixels) by 28.2 cm high (1024 pixels) display scope. The background of the display scope was white.

[3] The computer models were created by Viewpoint Datalabs. Many of the objects are available free on the World Wide Web (http://www.avalon.com). The others can be purchased from Viewpoint.

[4] The corresponding German names, which the subjects read before the experiment began, and which were displayed in the name verification experiments, were Flugzeug, Fahrrad, Auto, Stuhl, Klavier, Schuh, and Teekanne.

The display scope was viewed in a darkened room from a distance of approximately 100 cm, but the position of the observer's head was not fixed. The images ranged in size from 8.2 cm to 11.0 cm in diameter, so that the visual angle subtended by the images varied from 4.8 to 6.3°. The images were viewed binocularly, but without stereo. Naming responses were recorded by the computer to a digital sound file using a small microphone clipped to the observer's lapel. The sampling rate of the digital recording was 8000 samples per second at 8 bits, which yielded a 1 ms margin of error in the determination of the naming response time.

**Procedure.** The observers participated individually in 10-minute sessions. Each observer read printed instructions explaining that the task was to name a series of pictures of objects as quickly and as accurately as possible by saying the name of the object aloud into a microphone. The observer then completed three practice trials to ensure that the instructions had been understood. Following this, the observer read a list of the names of the objects (in German, see footnote 4) that would be seen. This was intended to reduce variability in naming times due to word finding difficulties. Each image was visible for 3 s. Presentation of the image on the display scope initiated a digital signal processing routine on the computer that recorded the observer's verbal response. The recording continued for the duration of the image. There was a 3-4 s interval following the image before the next trial began.

**Design.** The independent variables were the viewpoint from which the object was seen (View 1, View 2, or View 3) and the block in which each view was seen (Block 1, Block 2, or Block 3). The experiment was divided into three blocks of seven trials. Each object appeared once in each block, each time in a different view. Different objects were seen in different views within each block. To balance practice and repetition effects, the three views of each object were seen in different blocks by different observers.

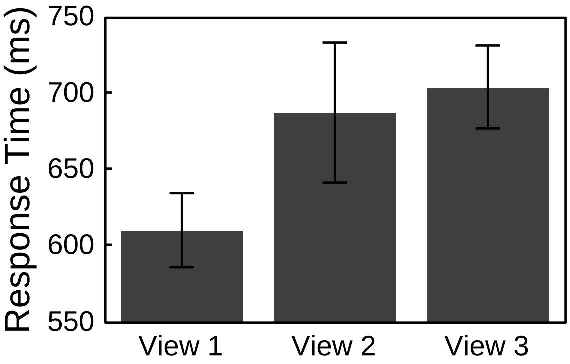

Figure 2: Mean naming response times for each view in Experiment 1a. Error bars indicate standard errors computed across observers.

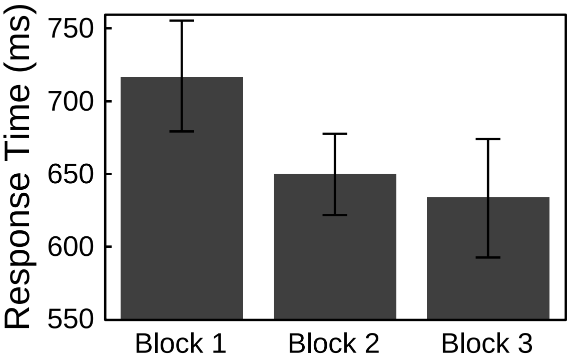

Figure 3: Mean naming response times for each block in Experiment 1a. Error bars indicate standard errors computed across observers.

## 1.2 Results and discussion

The accuracy of each response was determined by listening to the digital sound file that was recorded during the trial. Across all seven observers there were no naming errors. The response time for each trial was measured by locating the beginning of the response in the digital sound file with a thresholding routine.[5] Mean response times for each view collapsed over the three blocks and seven objects are presented in Figure 2. A one-way analysis of variance (ANOVA) with View as the independent variable revealed that the time to name the different views was significantly different

[5] Response times were also measured by visually examining the waveforms present in the sound files. Analyses of the data collected in this way were not qualitatively different from those performed on the data collected by the thresholding method.

[$F(2,12) = 6.900$, $MS_e = 2543.67$, $p < .05$]. A post-hoc analysis (Tukey's Honestly Significant Difference [HSD], $d_T = 72.0$ ms) revealed that naming times for View 1 were significantly faster than naming times for Views 2 and 3 ($p < .05$), but naming times for View 2 were not different from naming times for View 3 ($p > .05$).

Mean response times for each block collapsed over the three views and seven objects are presented in Figure 3. A one-way ANOVA with Block as the independent variable did not reveal a significant difference [$F(2,12) = 3.105$, $MS_e = 4361.34$, $p > .05$]. However, all seven observers named the objects faster in Block 1 than in Block 3 ($p < .01$). These results replicate the findings of Palmer et al. (1981) showing that subjectively preferred views are named more quickly than nonpreferred views, even when the main axis of the object is not appreciably foreshortened in the nonpreferred views. The additional finding that naming responses became faster after the first presentation (although the objects were seen in different views in subsequent presentations) was not reported by Palmer et al., but this finding is consistent with findings of other researchers indicating priming effects in object naming (e.g., Bartram, 1974; Biederman & Gerhardstein, 1993). Although it is not possible to determine whether the priming effects were verbally or visually mediated in the present experiment, the findings of Biederman and his colleagues (Biederman & Cooper, 1991a, 1991b, 1992; Biederman & Gerhardstein, 1993) suggest that at least some of the priming was visually mediated. It is also important to point out that the results of this experiment serve to validate the Spaceball technique developed by Blanz et al. (1995) for assessing preferred views.

## 2 Experiment 1b

With the knowledge that naming times vary for the different views studied in Experiment 1a, we are now in a position to assess whether performance will be different for these views in an objective task that does not involve object naming. In the following experiment, the same observers who participated in Experiment 1a performed a name verification task. On each trial the observer read the name of an object then saw a picture of an object and decided as quickly as possible whether the picture matched the name.

### 2.1 Method

The observers, stimuli, and apparatus were the same as in Experiment 1a. All observers completed Experiment 1a before beginning Experiment 1b. The observers responded by pressing the control key on the computer's keyboard. The names of the objects (see footnote 4) were displayed in black, Times Roman font. The height of each word was .72 degrees of visual angle, and the width varied, depending on the word, from 2.2 to 4.2°.

**Procedure.** Each observer participated individually in a 20-minute session. The observers read printed instructions explaining that their task was to decide as quickly as possible whether a picture of an object matched the name that was presented before the picture. They were informed that the names and pictures would be the same as those seen in the object-naming experiment they had just completed. On each trial one of the seven object names was displayed for 2-3 s in the middle of the display scope. The name was replaced by one of the three pictures of that object or by one of the pictures of a different object. The observer's task was to press the control key on the computer keyboard as quickly as possible if the picture matched the name or to do nothing if the picture did not match the name (a go/no-go task). The picture remained visible for 3 s or until a response was made, whichever was shorter. If the observer responded, the next trial began after a 1 s delay. If the observer did not respond within 3 s, a mismatch response was recorded and the next trial began after an additional 1 s delay.

**Design.** The three independent variables were the viewpoint from which the object was seen (View 1, View 2, or View

3), whether the picture matched the name (Match or Mismatch), and Block (Block 1—Block 6). Each block contained 42 match and 42 mismatch trials. For match trials, each of the seven names appeared with each of its three matching pictures twice (7 × 3 × 2 trials). For mismatch trials, each name appeared one time with one of the pictures of each mismatch object (7 × 6 trials). The viewpoints seen on mismatch trials were balanced so that each viewpoint was paired twice with each name, and each viewpoint of each object appeared twice. Over the entire experiment, each mismatch combination of name and picture appeared twice. These counterbalancing measures insured that each name was seen equally often in match and mismatch conditions and that each picture of each object was seen equally often in match and mismatch conditions. The order of the trials in each block was randomized differently for each observer. Before beginning the experiment, each observer completed 24 practice trials with the same three practice objects that had been seen in Experiment 1a. The name of each practice object appeared eight times, four times paired with its matching picture, and two times with each of the other mismatch pictures. The order of the 24 practice trials was randomized, but was the same for each observer.

## 2.2 Results and Discussion

Given the results of Experiment 1a, two questions of interest in this experiment were whether view canonicality would affect name verification time and whether name verification time would be influenced by repeated viewing of the same objects. Figure 4 shows the mean response times for each view collapsed over objects and blocks. Figure 5 shows the mean response times for each block collapsed over objects and views. Unlike in Experiment 1a, there were no systematic effects of viewpoint on response times, and observers did not respond more quickly following repeated exposure to the objects.

These observations were confirmed in a two-way View by Block within-subjects ANOVA, which revealed no significant differences for View [$F$(2,12) = 2.658, $MS_e$ = 840.01, $p > .05$], Block [$F$(5,30) = 1.403, $MS_e$ = 2119.87, $p > .05$], or the interaction between View and Block [$F$(10,60) = 1.676, $MS_e$ = 1379.16, $p > .05$]. Although the analysis presented above showed no significant effect of View or Block and no interaction between View and Block, it is possible that there was an effect of viewpoint very early in the experiment, and that this effect disappeared after the first exposure to each picture. We conducted two additional analyses to test this possibility. In the first analysis, we evaluated verification times for the *first* presentation of each view of each object. This analysis included only those trials in which a first presentation occurred on a match trial (recall that

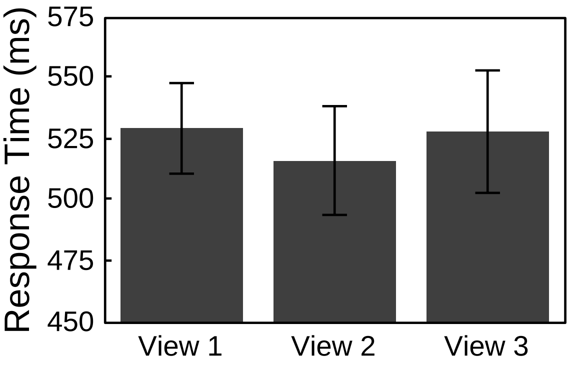

Figure 4: Mean naming response times for each view in Experiment 1b. Error bars indicate standard errors computed across observers.

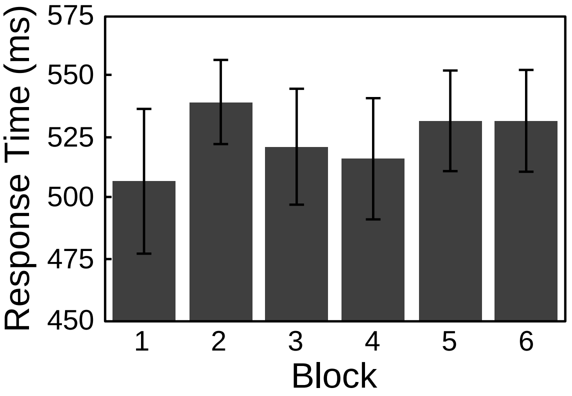

Figure 5: Mean naming response times for each block in Experiment 1b. Error bars indicate standard errors computed across observers.

subjects did not respond on mismatch trials, so we have response time data only for match trials). Excluding mismatch trials had the effect of reducing the number of trials used to compute the mean response time per view for each observer, but there was a sufficient number of responses to conduct the analysis. A one-way ANOVA with View as the independent variable revealed no significant differences [$F$(2,12) = 1.336, $MS_e$ = 3013.12, $p$ > .05]. The means (± standard error) were 454.2 (43.7) ms, 474.2 (54.6) ms, and 502.0 (46.1) ms for View 1, View 2, and View 3, respectively.

In the second first-presentation analysis, we evaluated verification times for the first *match* presentation of each view of each object. Thus, some responses included in this analysis were for pictures that had previously been seen in a mismatch trial. A one-way ANOVA again revealed no significant differences over View [$F$(2,12) < 1, $MS_e$ = 1784.46, $p$ > .05]. Mean response times (± standard error) were 505.8 (26.2) ms, 503.5 (34.3) ms, and 511.2 (37.3) ms for View 1, View 2, and View 3, respectively. There were few errors throughout the experiment. Summed over all seven observers and all blocks there were ten incorrect "match" responses for View 1, nine for View 2, and five for View 3. There was only one incorrect "mismatch" response for each view. The results of the present experiment contrast strongly with those of Experiment 1a. There was no evidence that view canonicality affected response times or accuracy in the name verification task. This was true even the first time each view was tested. Furthermore, there was no clear priming effect as there had been in Experiment 1a. Response times did not decrease in later blocks of the experiment.

One potential limitation with the present experiment, however, is that the observers' experience with the objects in Experiment 1a could have affected their performance. Having seen the objects in the naming task (one time in each view) could have been sufficient to eliminate any effects of view canonicality. Before we discuss the implications of this experiment, we will present a second name-verification experiment, nearly identical to the present experiment, using a new group of observers who were unfamiliar with the particular object models used in the experiment. To examine whether view canonicality affects the time needed to decide that a picture does not match a name, we used a two-response procedure in Experiment 2. Observers pressed one key to respond that the object matched the name and another key to respond that it did not match.

## 3 Experiment 2

### 3.1 Method

The stimuli, apparatus, and design were the same as in Experiment 1b.

**Observers**. Ten student volunteers from Eberhard-Karls University in Tübingen, Germany participated in the experiment for a payment of 10 DM. All were native speakers of German and reported having normal or corrected to normal visual acuity. None of the observers had participated in Experiment 1.

**Procedure**. The procedure was almost identical to that in Experiment 1b except that observers pressed one control key on the computer keyboard with their dominant hand to respond that the picture matched the name, and they pressed the other control key with their nondominant hand to respond that the picture did not match the name. Response feedback was provided in the form of a tone for incorrect responses. Although this two-response procedure increased response times overall, it allowed us to examine whether view canonicality affects the time needed to determine that a picture does not match a given name. The only other difference between this experiment and Experiment 1b was that the name was always displayed for 2 s rather than 2-3 s.

### 3.2 Results and Discussion

Mean correct response times for each view collapsed over all six blocks and seven objects are shown in Figure 6. Although responses were consistently faster for trials in which the picture matched the name, there was no systematic effect of view for either match or mismatch trials. Figure 7 shows the mean response times for each block collapsed over the three views and seven objects. There were small decreases in response times as the experiment progressed, both for match and mismatch trials.

These observations were confirmed in a three-way Match by View by Block within-subjects ANOVA. The ANOVA revealed that match trials were responded to more quickly than mismatch trials [$F(1,9) = 24.134$, $MS_e = 15347.06$, $p < .01$] and that response times were different across blocks [$F(5,45) = 2.949$, $MS_e = 7717.08$, $p < .05$]. Tukey's HSD indicated that responses were significantly faster in block 6 than in block 1 ($d_T = 47.8$ ms), but no other differences were significant. The main effect of View was not significant [$F(2,18) = 1.420$, $MS_e = 931.08$, $p > .05$], nor were any of the interactions among the variables.

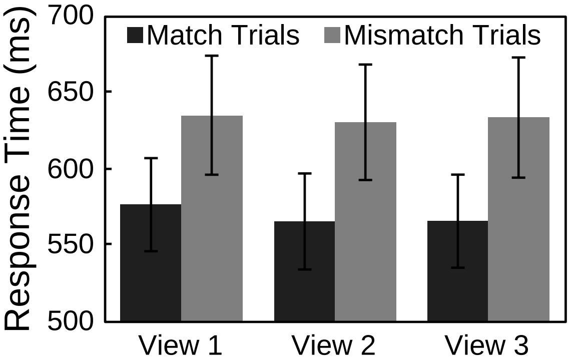

Figure 6: Mean response times for each view in Experiment 2. Dark bars indicate trials in which the name and the picture matched. Light bars indicate trials in which the name and the picture did not match. Error bars are standard errors computed across observers.

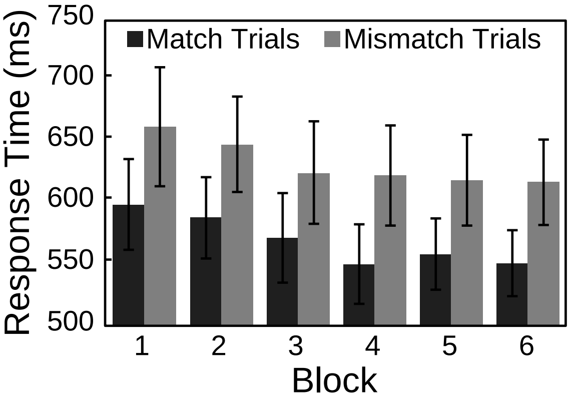

Figure 7: Mean response times for each block in Experiment 2. Dark bars indicate trials in which the name and the picture matched. Light bars indicate trials in which the name and the picture did not match. Error bars are standard errors computed across observers.

To examine whether there were effects of view canonicality early in the experiment, we carried out two analyses similar to those conducted in Experiment 1b. In the first analysis, we examined response times for the first presentation of each picture. These trials were divided according to whether they were match or mismatch trials. A two-way Match by View ANOVA revealed only a marginally significant advantage for match trials [$F(1,9) = 4.534$, $MS_e = 4690.38$, $p < .07$]. The main effect of View [$F(2,18) < 1$, $MS_e = 5540.36$, $p > .05$] and the interaction between Match and View [$F(2,18) = 1.062$, $MS_e = 8003.93$, $p > .05$] were not significant.

We conducted a second first-presentation analysis by examining response times for the first presentation of each picture in a match trial and the first presentation of each picture in a mismatch trial. A two-way Match by View ANOVA revealed that match trials were responded to more quickly than mismatch trials [$F(1,9) = 6.255$, $MS_e = 6229.44$, $p < .05$], but neither the main effect of View nor the interaction between Match and View was significant [both $F < 1$]. Although there were slightly more errors in this experiment than in Experiment 1b, they did not vary systematically with view. Over the entire experiment there were 8, 19, and 9 incorrect "match" responses and 19, 18, and 18 incorrect "mismatch" responses for Views 1, 2, and 3, respectively.

Response times were slower overall compared to Experiment 1b. The mean response time for correct trials in Experiment 1b was 524.8 ms, whereas mean response times in Experiment 2 were 568.2 ms for match trials and 632.3 ms for mismatch trials. The most likely explanation for this difference is the two-response procedure used in Experiment 2. The decrease in response times in later blocks is consistent with this explanation, and is likely due to the observers learning which hand to respond with on match and mismatch trials. Response times in the last block—550.8 ms for match trials and 617.1 ms for mismatch trials—were more similar to, but still somewhat longer than, those found in Experiment 1b. The difference in response times between match and mismatch trials might lead one to argue that visual or decision processes are somehow different for determining that a picture does not match a previously presented name. However, the most likely explanation for the difference in response times on match and mismatch trials is that the observers were simply quicker to respond with their dominant hands. To fully examine this issue it would be necessary to run an equal number of observers responding with their nondominant hands on match trials. However, we did not use the two-response procedure in the present experiment to examine differences in decision processes for match and mismatch trials. Rather, the purpose was to determine whether view canonicality would affect response times on negative trials. This was clearly not the case. If there is a different decision process being carried out on negative trials, the efficiency of the process does not appear to be affected by view canonicality.

As discussed above, Newell and Findlay (1996) also used a name verification procedure to study object identification. Unlike in the present study, they found significant differences in verification times for different views of objects. These differences, however, were related mostly to the degree of axis foreshortening. Verification times were roughly constant for views that were not appreciably foreshortened and were significantly longer only for views in which the object's main axis of elongation was rotated no more than 30° from the line of sight. Similar results were obtained by Lawson (1994) in an object naming task.

## 4 General Discussion

The purpose of this study was to examine view canonicality using two different objective tasks, object naming and name verification. As in previous studies, there were clear differences in the time needed to name different views of the objects. In Experiment 1a, observers named objects seen in preferred views more quickly than objects seen in nonpreferred views. No differences were found between preferred and nonpreferred views in Experiments 1b and 2 in which observers performed a name verification task. To examine the implications of these results, we will consider how they must be explained by each of the models of long-term visual memory that was presented in the introduction. We will then present an alternative explanation of canonicality effects based on view similarity.

### 4.1 Object-centered structural descriptions

Theorists proposing that an object's long-term visual representation consists of a single 3D object-centered structural description have argued that differences in the recognizability of different views of an object are due to differences in the ability to derive a structural description from those views (e.g., Marr, 1982). The results of the naming task presented here could be interpreted in this way, but the failure to find an effect of view in the name verification task poses a problem. If we accept the hypothesis that view effects arise in object naming because structural descriptions are difficult to derive from some views, then the lack of an effect of view in name verification must mean that it is not necessary to derive a complete structural description to perform the name verification task.

Marr (1982) touched on this issue in his discussion of how an object might be recognized from viewpoints in which a full description of the object can not be recovered. He speculated that in some cases it might be possible to recognize the object by constructing a partial description based on the axes that are visible from that view. Alternatively, recognition might be achieved by recognizing a particular part or only a few of the parts of an object. For example, an animal might be recognized in a frontal view, in which its main axis of elongation is completely foreshortened, by recognizing its face.

With regard to the present study, it could be that a complete recovery of structure is needed to name an object, which could be slow for some views, but a complete recovery is not needed to perform the name verification task. In particular, it might be necessary to recover only a few salient parts or features, rather than a complete description, to perform the name verification task. Recognition achieved in this way might be much less sensitive to viewpoint, as long as the necessary parts or features are visible in the image.

Recognizing an object in this way, however, would seem to be a substantial departure from the way in which recognition is believed to be achieved according to these theories. It is not even clear that the data needed to recognize partial descriptions of objects would be readily available from the object models stored in long-term visual memory. Allowing for recognition in this way might necessitate substantial changes to these theories.

It is important to consider that the need or opportunity to recognize objects using partial descriptions might be quite common. It often occurs that some of an object's parts are occluded from view either by the object itself or by other nearby objects. In these situations it would be necessary to recognize the object using an incomplete description or to complete the object by inferring what is occluded. Furthermore, the opportunity to recognize objects as in the name verification experiments presented here might be quite common. Consider, for example, a search task in which an observer must locate a particular object in a cluttered environment. This task, arguably a common visual task, is much more like a verification task than a naming task. Consider also the effect that scene context might have on object recognition. Scene context might set up expectations, not unlike the expectations set up in the name verification task, that minimize the need for a complete recovery of structure prior to recognition. These considerations draw into question the need to fully recover an object's 3D structure, especially given the high cost of recovering object-centered 3D structure from a 2D image.

## 4.2 Viewpoint-dependent descriptions

In many models of long-term visual memory that represent objects with viewpoint-dependent descriptions (whether 2D or 3D), it is argued that differences in the speed of recognition of different views of an object arise because an error prone or time consuming transformation mechanism must be used to recognize views for which no description is explicitly represented. For example, these views might be recognized by interpolating among views that are represented (e.g., Poggio & Edelman, 1990) or by transforming either the input view, the stored views, or both (e.g., by mental rotation, Shepard & Cooper, 1982, Tarr, 1995, or by alignment, Ullman, 1989, 1996).

For reasons similar to those given above in the discussion of object-centered models, the results of the present study are not in accord with the predictions of these view-dependent models. If transformation mechanisms were responsible for the differences in naming speed observed in Experiment 1a, then we must conclude that these mechanisms were not used to perform the name verification task in Experiments 1b and 2. This again begs the question of why such mechanisms would ever be used. Some researchers have argued that transformation mechanisms are not normally used to recognize objects, but that they might be used to confirm recognition decisions or to make decisions about object properties that depend on an external reference frame, for example, handedness judgments (Corballis, Zbrodoff, Shetzer, & Butler, 1978; Corballis, 1988). It is not at all clear, however, why object naming would require confirmatory processing and name verification would not. Furthermore, there appears to be no reason why naming would involve external reference frames and name verification not. In fact, it is object naming that is generally considered to be the task not requiring such additional processing.

Perrett et al. (1996) present an alternative view-based model of long-term visual memory that does not rely on view transformation mechanisms. According to their model, differences in the speed of recognition of different views of an object arise because different views are represented with different weights. In particular, they argued that different numbers of neurons are recruited to encode different views of an object depending on the observer's experience with those views. Views that are experienced often are represented by a greater number of neurons than views experienced less often. The result is that evidence regarding the identity of a viewed object will accumulate more slowly for views that are either weakly represented or not at all explicitly represented in comparison to views that are more strongly represented. At first sight, the results of the present study seem to be at odds with this theory. The differences in naming speed observed in Experiment 1a would suggest that View 1 is more strongly represented than Views 2 and 3. Why then does this not affect the speed of recognition in Experiments 1b and 2? One explanation might be that less evidence is needed to perform the verification task than the naming task. Perrett et al. showed that at low thresholds the time needed to accumulate sufficient evidence is more similar for strongly and weakly represented views. Although these claims are only speculative, the possibilities are intriguing and should be considered in future research.

### 4.3 View similarity

The need to resort to some alternate mode of recognition to explain differences in view effects in different tasks is, at best, unsatisfactory. What is needed is a model of the recognition process that can account for the kinds of differences observed in the present study without having to alter the basic mechanism by which recognition is achieved. In this section we provide a sketch of one such model, which bases recognition on the similarity between an input view and relevant views stored in memory.

Assume that, as in the multiple-view theories discussed above, an object's long-term visual representation consists of a collection of more or less viewpoint-specific descriptions. Figure 8 depicts multiple-view representations of two familiar objects, a camel and a giraffe. To recognize an object seen in a particular view, the internal description of that view is compared to descriptions of views stored in memory, and the object is recognized as a member of the object class to which the best matching, or most similar, description belongs. We will assume that there are a sufficient number of descriptions of each object stored so that it is possible, in principal, to recognize familiar objects from nearly any viewpoint.

Consider what might happen if a view such as the frontal view of the camel depicted in Figure 8 had to be identified. In the absence of contextual cues or prior hypotheses about what object might be seen, it would be necessary to compare the description of this view with a large number of stored descriptions. The description of this frontal view is likely to be sufficiently similar to stored descriptions of more than one object. In the example depicted in Figure 8, the frontal view of the camel is seen to be similar to the frontal view of the giraffe as well. Similarity to views of more than one object class could lead to response competition, which could slow identification of the object in this view. A similar effect could occur for the view of the back of the camel shown in Figure 8.

On the contrary, the profile or three-quarter views of the camel are not likely to be confused with views of other objects. These views contain distinctive features, for example, the hump on the camel's back, that minimize the possibility that they will be con-

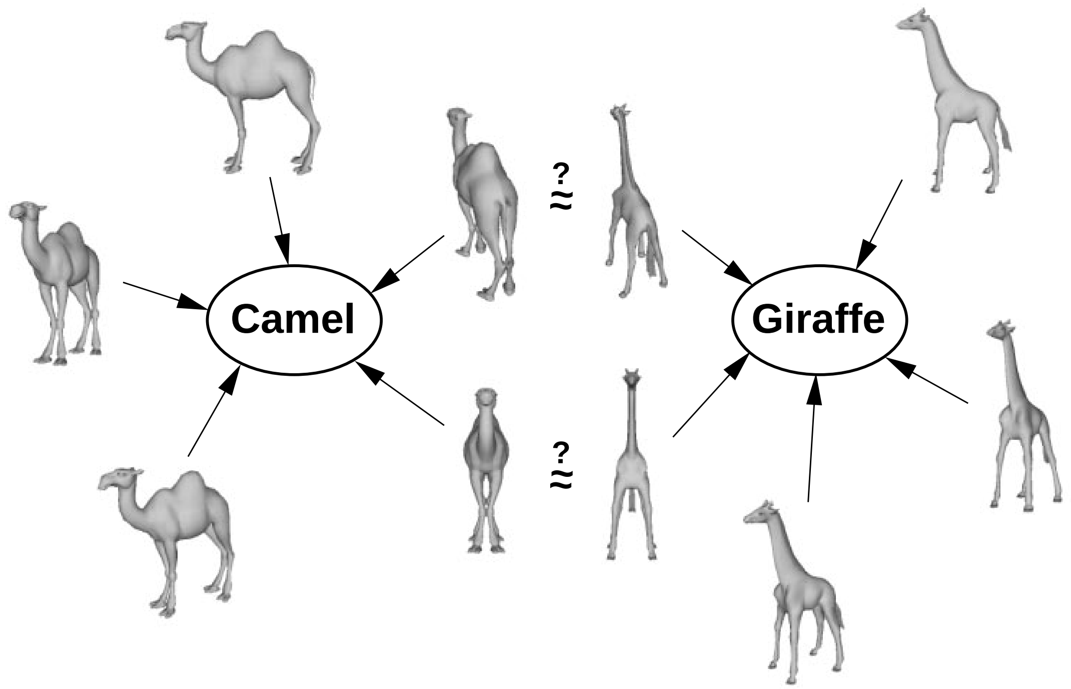

Figure 8: Multiple-view representations of two objects. Notice that some views could be confused, which might lead to response competition in a bottom-up task such as object naming. There are other views, however, that are sufficiently distinct so as not to be confused.

fused with views of another object, even if they must be compared to a large number of views of other objects. Recognizing an object in such a view would be faster relative to recognizing the object in a view that is similar to views of more than one object, because the likelihood of response competition is reduced.

Why then does view similarity not affect performance in the name verification task? One simple explanation is that providing an explicit hypothesis regarding the object's identity reduces the space of views to which the input view must be compared. This need not, in itself, reduce the time needed to make a recognition decision.[6] All that is necessary is that it reduce the likelihood that the input view is confused with views of other objects. In the present study it is unlikely that any of the tested views would have been easily confused as views of other objects in the set. According to the view similarity model presented here, this would have eliminated differences in verification times for the different views. Notice that in the naming task, one cannot consider only the other objects in the set when assessing the likelihood of confusions. Confusions could arise in naming from any number of objects that the observer has previously experienced.

The arguments presented above suggest that view canonicality is not simply a function of the task, but is instead related to the space of objects and views that must be considered to perform a particular task. This suggests that view effects might be found in a verification task that included highly similar objects, for example, the camel and the giraffe shown in Figure 8. Deciding whether the frontal view of the camel was a giraffe might be slow and sometimes incorrect, but making the same de-

[6] We are not suggesting that views are compared in a serial fashion so that the number of views that must be compared affects the time needed to make a recognition decision. View "comparisons" could be made in parallel in a neural network so that little or no additional cost is incurred by simply increasing the number of views that must be compared.

cision for the profile or three-quarter view of the camel might be fast and accurate. This is a topic for further research that should begin with a careful determination of inter-object similarity.

## References

Bartram, D. J. (1974). The role of visual and semantic codes in object naming. *Cognitive Psychology*, **6**, 325–356.

Bartram, D. J. (1976). Levels of coding in picture-picture comparison tasks. *Memory & Cognition*, **4**, 593–602.

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., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, **18**, 121–133.

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., Vetter, T., Bülthoff, H. H., & Tarr, M. J. (1995). What object attributes determine canonical views? *Perception*, **24**(Supplement), 119c.

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.

Corballis, M. C. (1988). Recognition of disoriented shapes. *Psychological Review*, **95**(1), 115–123.

Corballis, M. C., Zbrodoff, N. J., Shetzer, L. I., & Butler, P. B. (1978). Decisions about identity and orientation of rotated letters and digits. *Memory & Cognition*, **6**(2), 98–107.

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.

Ellis, R., & Allport, D. A. (1986). Multiple levels of representation for visual objects: A behavioural study. In A. G. Cohn & J. R. Thomas (Eds.), *Artificial Intelligence and its Applications* (pp. 245–257). New York, NY: J. Wiley.

Harries, M. H., Perrett, D. I., & Lavender, A. (1991). Preferential inspection of views of 3-D model heads. *Perception*, **20**(5), 669–80.

Humphrey, G. K., & Jolicoeur, P. (1993). An examination of the effects of axis foreshortening, monocular depth cues, and visual field on object identification. *The Quarterly Journal of Experimental Psychology*, **46A**(1), 137–159.

Lawson, R. (1994). *The effects of viewpoint on object recognition.* Unpublished Ph.D. Thesis, University of Birmingham, School of Psychology.

Liter, J. C. (1996). The influence of qualitative and quantitative features on object recognition across changes of view. Manuscript submitted for publication.

Lowe, D. (1985). *Perceptual Organization and Visual Recognition.* Boston: Kluwer.

Lowe, D. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, **31**, 355–395.

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. (1996). The effect of depth rotation on object identification. Manuscript submitted for publication.

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

Perrett, D. I., & Harries, M. H. (1988). Characteristic views and the visual inspection of simple faceted and smooth objects: 'tetrahedra and potatoes'. *Perception*, **17**(6), 703–20.

Perrett, D. I., Harries, M. H., & Looker, S. (1992). Use of preferential inspection to define the viewing sphere and characteristic views of an arbitrary machined tool part. *Perception*, **21**(4), 497–515.

Perrett, D. I., Oram, M. W., & Wachsmuth, E. (1996). Evidence accumulation in cell populations responsive to faces: An account of generalisation of recognition without mental transformations. Manuscript submitted for publication.

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**(4), 197–208.

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

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations.* Cambridge, MA: MIT Press.

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

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, **32**, 193–254.

Ullman, S. (1996). *High Level Vision.* Cambridge, MA: MIT Press.