Generalization between canonical and non-canonical views in object recognition

Tandra Ghose
Department of Psychology,
Technical University of Kaiserslautern, Germany

Zili Liu
Department of Psychology,
University of California Los Angeles, USA

Viewpoint generalization in object recognition is the process that allows recognition of a given 3D object from many different viewpoints despite variations in its 2D projections. We used the canonical view effects as a foundation to empirically test the validity of a major theory in object recognition, the view-approximation model (Poggio & Edelman, 1990). This model predicts that generalization should be better when an object is first seen from a non-canonical view and *then* a canonical view than when seen in the reversed order. We also manipulated object similarity to study the degree to which this view generalization was constrained by shape details and task instructions (object vs. image recognition). Old-new recognition performance for basic and subordinate level objects was measured in separate blocks. We found that for object recognition, view generalization between canonical and non-canonical views was comparable for basic level objects. For subordinate level objects, recognition performance was more accurate from non-canonical to canonical views than the other way around. When the task was changed from object recognition to image recognition, the pattern of the results reversed. Interestingly, participants responded "old" to "new" images of "old" objects with a substantially higher rate than to "new" objects, despite instructions to the contrary, thereby indicating involuntary view generalization. Our empirical findings are incompatible with the prediction of the view-approximation theory, and argue against the hypothesis that views are stored independently.

Keywords: view-based object recognition, canonical view, object categories

Citation: Ghose, T., & Liu, Z. (2013). Generalization between canonical and non-canonical views in object recognition. *Journal of Vision, 13*(1):1, 1–15, http://www.journalofvision.org/content/13/1/1, doi:10.1167/13.1.1.

Introduction

Human visual recognition of natural objects appears effortless. However, this apparent effortlessness is deceptive, because the visual system has to solve a difficult mapping problem in order to recognize an object. The problem is difficult because a three-dimensional (3D) object can give rise to a great number of possible 2D images due to variations of the viewing position, the illumination, and the scene surrounding the object. Among these variations, viewpoint variation is one of the most studied and most challenging, probably because an object can appear drastically different from one view to another, in part due to self-occlusion; for reviews, see Tarr and Bülthoff (1995); Logothetis and Sheinberg (1996); Ullman (1996); Edelman (1998); Biederman (2000); Hayward (2003); Kersten and Yuille (2003); Kersten, Mamassian, and Yuille (2004); and Palmeri and Gauthier (2004).

In everyday life, self-occlusion of objects is unavoidable because surfaces of an opaque 3D object closer to the observer will block light rays from farther surfaces. Such occlusion leads to a situation where parts of the object that are hidden from one viewpoint become visible from a different viewpoint. This situation poses a challenge for the recognition system to categorize the two different images as belonging to the same object. Object recognition theories propose various ways to categorize an incoming image as belonging to a specific category. In this study we summarized relevant theories of object recognition, developed a hypothesis based on the view-approximation model (Poggio & Edelman, 1990), observed human performance for viewpoint generalization, and tested the validity of this model.

The exact nature of the internal representation of objects in our visual system and the process of recognition when an object is encountered from a new viewpoint is an open question and a subject of prolonged debate (Biederman & Gerhardstein, 1993; Biederman & Gerhardstein, 1995; Tjan & Legge, 1998; Tarr & Bülthoff, 1995). There are two sets of theories to address this question.

Theory 1 The first theory hypothesizes that an object's 3D structure is represented in a viewpoint invariant manner, such that when a new view is shown, the visual system in effect attempts

doi: 10.1167/13.1.1 Received November 14, 2011; published January 8, 2013 ISSN 1534-7362

to find the appropriate 2D projection from this 3D structure in order to match the 2D incoming image (Binford, 1971; Marr & Nishihara, 1978).

There are three characteristics of this theory that are noteworthy.

a. The theory proposed as such is an ideal way to represent an object in memory. That is to say, if we ignore the context in which the representation is initially learned and the specific recognition task at hand, then fully encoding the 3D structure of an object is the best way to accomplish shape-based recognition of the object (Grenander, 1993). It is an empirical question however whether the visual system is capable of accomplishing this. For example, a variation of this theory focuses on encoding qualitative shape information while de-emphasizing metric information. This allows the qualitative shape description to be valid within a wide range of viewpoint variations (Biederman, 1987; Lowe, 1987).

b. This theory does not take into consideration how a representation is acquired, and how the viewing history of an object may influence recognition of the object. To be fair, however, this is probably so not because these factors were considered unimportant, but because the theory is primarily focused on what an established representation in long-term memory ideally should be. Nevertheless, if the viewing history is not considered, problems may arise when testing this viewpoint invariant theory. For example, a common practice in experimentally testing this theory is to first present an object from viewpoint A. Subsequently, recognition is tested when the object is shown either from viewpoint A, or from a different viewpoint B (Tarr, Williams, Hayward & Gauthier, 1998). The hypothesis is that comparable performance between A and B supports the viewpoint invariant theory. Otherwise, it suggests that the representation is viewpoint dependent. The empirical results, across a large range of tasks and stimuli, are that recognition performance is better for viewpoint A than for viewpoint B. However, it takes time for the visual system to update the representation with the recently acquired information from viewpoint A. In other words, the "propagation" of the new information from a viewpoint specific format into a less specific format is not immediate. When these dynamics are considered, disentangling the representation in longer-term memory becomes complex. We also take issue with the assumption that if the representation is viewpoint invariant, recognition is necessarily comparable between viewpoints A and B. This will be elaborated below.

c. Although this theory emphasizes the viewpoint independent nature of the representation, it is possible to modify the theory to allow the representation to be viewpoint dependent, for the following reason. There is an ecological advantage to representing different viewpoints unequally; for example, a person is more likely to be viewed at eye level than from top down. Accordingly, it is sensible to distribute representational accuracy differently to better use limited resources.

Theory 2. The second set of theories proposes that the memory representation is viewpoint dependent, either because the visual system is computationally incapable of building a viewpoint invariant representation or because it is sensible to deliberately encode the viewpoint from which an object was seen (Poggio, 1990, Tarr & Bülthoff, 1998). A key feature of these theories is that the representation is a collection of previously seen views of an object, but there is no attempt to reconstruct its 3D structure. Specifically, the fact that different views of a rigid object are related is not represented (including, for example, the fact that the object can be rotated from one view to the next). In other words, the constraint that different views form into a geometrically consistent 3D structure is missing.

We now use two examples to illustrate how an incoming view may be matched to a set of 2D previously seen views. The first example involves a linear combination of some of the 2D images from the stored set, and the second example involves no combination at all. Assuming that an object's view is represented by the 2D coordinates of its features, Ullman and Basri (1989, 1991) showed that any view of an object is a linear combination of two other views of the same object (the features are assumed visible from all views). In other words, if a new view cannot be represented by a linear combination of two views of an

object, then this new view is not projected from this object. Alternatively, in a view-approximation model (Poggio & Edelman, 1990), the previously seen views are stored independently in memory. When a new view is seen, recognition is achieved by comparing a similarity measure between this new view with each stored view, and the summation of the total similarity determines whether the new view is recognized or rejected. Whether this new view can be expressed as a linear combination of stored views is not considered.

How might these two sets of theories be tested empirically? During the prolonged debate in the past three decades, scientists have focused on the following criterion in regard to whether the internal representation is viewpoint dependent or independent (Binford, 1971; Marr & Nishihara, 1978; Biederman, 1987; Tarr & Pinker, 1989; Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992); see also Riesenhuber and Poggio (1999, 2000) and Hayward (2003) for a review. If recognition performance (measured in accuracy or speed) varies as a function of viewpoint, then the representation is viewpoint dependent. This assumption, commonly held by all in the debate, is mistaken, however, because object recognition necessarily requires matching between an incoming view with the object's internal representation (Marr, 1982; Liu, 1996; Liu, Knill & Kersten, 1995). Therefore, even if the representation is viewpoint invariant and all viewpoints are assumed equally likely, a more informative view should in theory give rise to better recognition performance than a less informative view. The well-known accidental view effect illustrates this point. For example, the observation that a bucket is hard to recognize from the top-down view (Warrington & Taylor, 1973) may be attributed to the impoverished stimulus information, even if the representation is viewpoint invariant. The well-known canonical view effect illustrates the point from the opposite end (Palmer, Rosch & Chase, 1981; Blanz, Tarr, Bülthoff, & Vetter, 1999).

A canonical view is an image of an object that is the most representative, comes to mind first when associating a name, and gives rise to the most accurate and fastest recognition performance. For example, a three-quarter view of a horse is a canonical view of a horse. Since the original discovery of canonical views by Palmer et al. (1981), a large number of studies have confirmed that canonical views indeed gave rise to superior recognition performance relative to non-canonical views (Perrett & Harries, 1988; Harries, Perrett, & Lavender, 1991; Edelman & Bülthoff, 1992; Perrett, Harries, & Looker, 1992; Cutzu & Edelman, 1994; Bulthoff, Edelman, & Tarr, 1995; Verfaillie & Boutsen, 1995; Newell, Ernst, Tjan, & Bülthoff, 2001). This finding led to the hypothesis that canonical views were encoded into a viewpoint-dependent representation. However, this conjecture is premature because, as argued above, the superior performance can be attributed to the representation only if a canonical and a non-canonical view are assumed to be equally informative. Gomez, Shutter, and Rouder (2008) made exactly this assumption *that canonical and non-canonical views were equally informative and that the canonical view effect was due to viewing frequency*. They further conjectured, based on their empirical results, that the canonical view effect was analogous to the word frequency effect in word memory tests. Essentially, they argued that canonical view effects could be completely accounted for by the possibility that a canonical view of an object is more frequently seen than a non-canonical view. However, in this study we will argue for an alternative perspective.

As explained above, behavioral performance on object recognition tasks is dependent on both internal representation and the information content of the incoming view. It should now be apparent that, although the canonical view effects are well accepted and robust, inferring the nature of the internal object representation from these effects is difficult. In this study, we set out to test a specific version of the second set of theories about the internal representation of objects: the view-approximation model by Poggio and Edelman (1990).

The view-approximation model

A specific and prominent viewpoint dependent theory in object recognition is termed view-approximation model (Poggio, 1990; Poggio & Edelman, 1990; Edelman & Poggio, 1992). Although there are several variations of the model, its central theme is that the representation R of an object is a set of previously seen views V_i: where $R = \{V_i, i = 1 \ldots n\}$. When an object is seen, its input view V is compared with each stored "view" V_i to yield a similarity measure $s(V, V_i)$. A summation of these pair-wise similarities gives rise to an overall measure of similarity $s(V, R) = \sum_i s(V, V_i)$. Consequently, the input object is either recognized as the same, or "old," object represented by R if the overall similarity is above a certain threshold; or otherwise rejected as a different, or "new," object.

We now use the classic results on canonical views to make a prediction from the model above, a prediction that has never been discussed in the literature, as far as we know. The classic results claim that a canonical view gives rise to more accurate object recognition than a non-canonical view. Accordingly, a straightforward prediction from the model above is as follows. We first assumed that, after seeing a view V of an object, the object representation R is updated to be $V \cup R$ the prediction is: if an object is first studied from a canonical view V_{cc} and then tested from a non-

canonical view V_{nc}, recognition performance for the most recent view will be worse than if the view sequence is reversed. Namely,

$$s(V_{nc}, V_{cc} \cup R) - s(V_{cc}, V_{nc} \cup R) = s(V_{nc}, R) - s(V_{cc}, R) < 0. \quad (1)$$

Note that this inequality holds irrespective of how the similarity $s(\bullet)$ is defined.

Our justification for using the union of V_{cc} and R as the updated representation after viewing V_{cc} is that even for a canonical view V_{cc}, this exact view from this exact synthetic object is unlikely to have been seen by participants. This relationship is consistent with the model proposed by Poggio and Edelman (1990). Namely, even if a canonical view has been seen before, adding it again to the representation R is equivalent to increasing the coefficient from this view to the cumulative similarity measure.

Equation 1 also holds in the following two scenarios that are particularly relevant to the experiments in this paper:

1) Scenario 1: the similarity comparison takes into consideration all stored views, as Equation 1 indicates. This consideration is in a situation when an object needs to be recognized regardless of its viewpoint (that is, in an object-recognition task). Then according to Bayesian inference (Knill & Richards, 1996), all stored views should be considered. This integration can be further generalized. For example, the stored views do not have to be equally weighted. The views that are more informative or more frequently seen can be weighted more. These generalizations, however, do not change the inequality in Equation 1.
2) Scenario 2: the similarity comparison does not take into consideration all stored views. Instead, a nearest neighbor operation, which considers only one stored view that best matches the input view, may be used. This operation may be a reasonable alternative when both an object and its original viewpoint need to be identified: for example, in situations where one may need to perform image recognition task, that is, whether the exact image of a given object had been seen before. The question now is whether Equation 1 would still hold.

In order for Equation 1 to hold, the stored view in the representation that is closest to the canonical view needs to give rise to a higher similarity than its non-canonical counterpart. The following two possibilities suggest that Equation 1 still holds.

2.1) The first possibility is that the distribution of viewpoints across the viewing sphere is non-uniform, such that canonical views are more likely to be stored than non-canonical views. For example, a top-down view of a human is less likely than views from the sides. In this case, as Gomez et al. (2008) argued, based on mere frequency of occurrence, on the viewing sphere a stored view is likely to be closer to the canonical than to the non-canonical view.

2.2) The second possibility is that the nearest stored view is likely to be more similar to a canonical than to a non-canonical incoming view, even if the viewpoint distribution over the viewing sphere is uniform. This is because views that are qualitatively similar are clustered over the viewing sphere (Koenderink & van Doorn, 1976). As an example, imagine the viewing sphere of a cube in parallel projection. The image of a cube shows either one, two, or three sides of the cube, with increasing surface cluster area on the viewing sphere. If one accepts the assumption that a canonical view of a cube has three, but not one or two, sides visible, then a randomly selected view on the viewing sphere is more likely to be more similar to a canonical than to a non-canonical view.

Weinshall and Werman (1997) provided further mathematical support to this possibility, although it remains an open question whether their mathematically defined "typical," "characteristic," or "generic" views are synonymous to the psychological canonical views. It should also be noted that Weinshall and Werman (1997) used generic notions of similarity. In fact, if views are assumed to be independently stored, then Weinshall and Werman (1997) also predict Equation 1 in Scenario 1. No prediction can be deduced from Weinshall and Werman (1997), however, if views are not assumed to be stored independently.

Another important yet unresolved question is how the representation of an object is updated. What we assumed here is perhaps one of the simplest postulations that follows from Poggio and Edelman (1990): for Equation 1, the assumption is that the new view in the study phase changes the representation of the object in the following sense. Denoting the new view as V_n and the representation as R, then the changed representation is $V_n \cup R$. Namely, V_n is added as an additional item to the set of stored views represented as R. This assumption is made by the following equation in Poggio and Edelman (1990):

$$f(x) = \sum_{\alpha=1}^{K} c_\alpha G(||x - t_\alpha||),$$

where t_α ($\alpha = 1, \ldots, K$) is a stored view, x is an input view, and G is a radial basis function (or a Gaussian in the implementation), c_α is the coefficient, and $f(x)$ is a measure of the overall similarity between the input x and the object represented by t_α ($\alpha = 1, \ldots, K$). There is

no principled way to determine the coefficient c_α, which is assumed to be independent of α when all views are treated equally. Apparently, all views are stored independently of each other and contribute independently to the overall similarity measure.

A modern relevance of this assumption is illustrated in Edelman and Shahbazi (2012)

$$CT(x) = \frac{1}{\sqrt{n}} \begin{pmatrix} ||x - p_1|| \\ \vdots \\ ||x - p_n|| \end{pmatrix}.$$

Aside from the notation changes, it is apparent that the independent assumption of the stored "prototypes" and their role in contributing to the overall similarity is fundamentally the same as in Poggio and Edelman (1990).

To summarize, we used the canonical view effect as a foundation to investigate the validity of a major theory in object recognition, the view-approximation model. We compared situations where a familiar object was first studied from a canonical view, and then tested from a non-canonical; and vice versa. We manipulated the recognition task such that the required recognition was either viewpoint irrelevant (Scenario 1) or relevant (Scenario 2). The arguments above predicted that Equation 1 would hold in both tasks, if views were assumed to be stored independently. In order to increase the power of our hypothesis testing, we parametrically manipulated our experiments by testing object recognition either in the basic level category or in the subordinate level category (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). For example, "bird" would be a basic level categorization and "robin," "sparrow," "jays," and "crows" would be exemplars that belong to the subordinate level of categorization. By extension, we expected the differential effect as calculated in Equation 1 to be more pronounced when the task was changed from between-category (e.g., a cow vs. a car) to within-category object recognition (e.g., a BMW vs. a Mercedes). Our rationale was that, when within-category distractors were added, additional view-specific shape details would be required for successful object recognition. At the same time, Equation 1 was expected to hold in all four conditions: (basic level, subordinate level) × (view irrelevant, view relevant recognition). We tested this pattern of predictions.

To anticipate the results, we found that Equation 1 held only in certain experimental conditions, but was violated in other conditions. These findings contradicted the view-approximation theory that predicted that Equation 1 should hold in all conditions. Our results argued against the hypothesis that views are stored independently. In what follows, we will present three experiments. Experiment 1 served to validate that our choice of a canonical view was more representative of an object than the non-canonical view. Experiment 2 studied viewpoint irrelevant object recognition (Scenario 1) in basic and subordinate level categories, and the results were consistent with Equation 1. Experiment 3 was nearly identical to Experiment 2, except viewpoint relevant image recognition (Scenario 2) was studied, and the results were inconsistent with Equation 1.

Experiments

Experiment 1: Goodness-of-view rating

Palmer et al. (1981) used a set of preselected images of an object from different viewpoints and defined the canonical image as the one that was preferred over other images by participants. Using a similar paradigm, we preselected two images per object to verify that the canonical and non-canonical views chosen by the experimenters were agreed upon by naïve participants.

Stimuli

Author TG chose a canonical (CC) and a non-canonical view (NC) from each of 144 computer graphics 3D objects. These images are available at this website: <http://www.sowi.uni-kl.de/fileadmin/wpsy/public/MOR/CNC.htm>. For most of the objects the canonical and non-canonical views differed by approximately 45 degrees in rotation around the y-axis. A canonical view was chosen such that many of the distinctive features of an object were visible. The non-canonical view was selected by rotating the object away from the canonical view such that at least one of the distinctive features (for example, a corner or a limb) became occluded. Among the 144 objects, 108 were basic level and 36 were subordinate level objects (12 cars, 12 dinosaurs, and 12 airplanes). The 36 subordinate level objects and 36 of the 108 basic level objects were from the Inventor object database (Silicon Graphics, Inc., Mountain View, California, version 2.0, 1992). The remaining 72 basic level objects were from the Tarr object database (http://www.tarrlab.org/). The images were grayscale, and rendered under orthographic projection with Lambertian shading. Figure 1 shows some example objects.

Apparatus

Stimuli were presented in a dark room on a 16″ calibrated computer monitor (Mitsubishi Diamond Plus 73) with a refresh rate at 85 Hz, and a resolution of 1024 × 768 pixels. The size of each image was 450 × 450 pixels, subtending a 14° × 14° visual angle at a

viewing distance of 57 cm. The experimental program was written in MatLab and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

Procedure

Two preselected images of the same object, rendered from two different viewpoints, were shown side by side for up to 5 s of viewing time, with their relative positions randomized. Participants decided which view was more representative of the object, by a mouse-click. After the response, the images were replaced by uniform gray. The participant pressed a mouse key to start the next trial.

Participants

Thirty-two undergraduate students of the University of California Los Angeles (UCLA) participated for course credits. All participants had normal or corrected-to-normal visual acuity, and were naïve to the purpose of the experiment.

Results

The canonical views were chosen 69.15% of the time as the more representative view, significantly higher than the 50% chance level ($t(31) = 10.72$, $p < 0.0001$) (Figure 1). This percentage was further broken down by object categories as follows: cars, 84.38%, $t(31) = 9.86$, $p < 0.0001$; dinosaurs, 73.44%, $t(31) = 6.94$, $p < 0.0001$; airplanes, 78.65%, $t(31) = 8.46$, $p < 0.0001$; miscellaneous items from the SGI Inventor database, 74.91%, $t(31) = 9.27$, $p < 0.0001$; and miscellaneous items from the Tarr object database, 62.72%, $t(31) = 4.90$, $p < 0.0001$. These results indicate that there was good consensus among participants that the canonical views chosen by the experimenter were more representative of the objects than were their non-canonical counterparts.

Experiment 2: Viewpoint irrelevant recognition of 3D objects

In this experiment participants performed an object-recognition task when objects were studied either from a canonical (CC) or a non-canonical view (NC) and then tested with a matched or mismatched view. This was done for both basic level and subordinate level objects.

Stimuli

The stimuli were the same images of the 144 objects in Experiment 1.

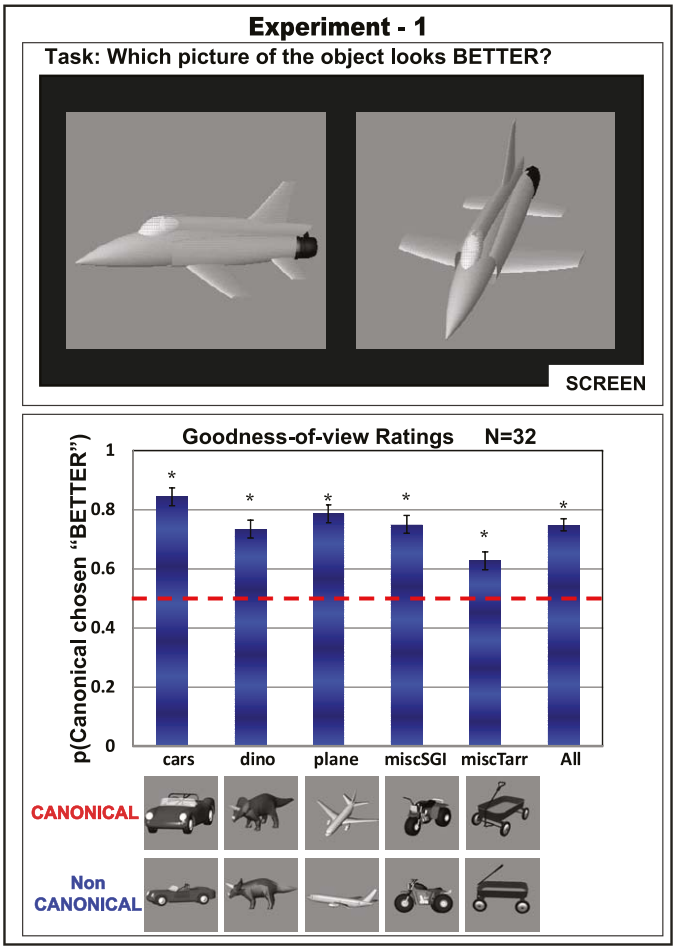

Figure 1. Experiment 1: Goodness-of-view rating. The purpose of this experiment was to see whether the view chosen as "canonical" by the experimenter was indeed preferred as a "better" view as compared to the non-canonical view of the same object. The participants chose the canonical as the better view significantly more often than at chance (the red-dashed line), for all classes of objects used. At the bottom are examples of canonical and non-canonical images of some objects.

Procedure

The experiment was an old-new rating study often used in memory research. There were two blocks, counter-balanced across participants. One block had only basic level objects, and the other block had subordinate level objects (12 cars, 12 dinosaurs, and 12 airplanes). Each block had two phases: study and test (Figure 2).

In the basic level block, the 108 objects were randomly divided into two halves. One half, consisting of 54 objects, was called "old" and was shown in the study, with half of them in canonical views (CC) and the other half in non-canonical views (NC). In the test, half of the "old" objects were shown identically as in the study. The other half's viewpoints were changed between canonical and non-canonical. The 54 "new"

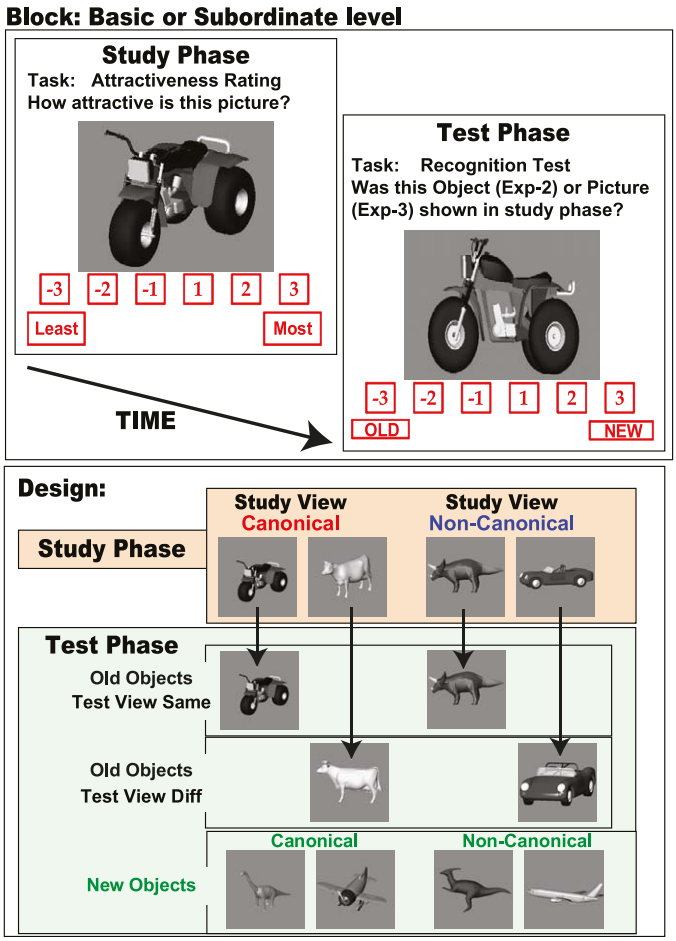

Figure 2. Experiments 2 and 3: Basic (108 objects) and subordinate level objects (12 cars, 12 dinosaurs, and 12 airplanes) were run in two separate blocks. Within each block a study phase with attractiveness rating task was followed by a test phase with old-new memory rating task. The memory rating task was on viewpoint-irrelevant objects (Experiment 2) or on viewpoint specific images (Experiment 3). In the study phase, an object was shown either in its canonical or non-canonical view. Subjects rated its attractiveness. In the test phase, either an "old" object was shown either in same/matched or different/mismatched view. An equal number of "new" objects were shown in either canonical or non-canonical view. No object was tested twice.

objects were shown only in the test phase, half of which were in canonical and the other half in non-canonical views. In the subordinate level block, the same procedure was followed, and half objects in each category (cars, dinosaurs, and planes) were "old" and half were "new."

More specifically, in the study phase, an object image was shown for 1 s and was replaced by a scale of (−3 −2 −1 +1 +2 +3). The following words were written below their corresponding numbers in the scale: (least attractive, −3), (below average, −1), (above average, +1), and (most attractive, +3). Participants responded by selecting a number, and the next trial started automatically. The subordinate level block was preceded by a reminder that there was a high degree of resemblance between some objects and that care was needed to do well. Each subordinate level object was shown twice in the study, as compared to the basic level block where each studied object was shown only once.

In the test phase, an object image was shown for 1 s, and was replaced by a scale of (−3 −2 −1 +1 +2 +3). For half of the participants, the following words were written below their corresponding numbers in the scale: ("surely old," −3), ("guess old," −1), ("guess new," +1), and ("surely new," +3). For the remaining participants, the old-new direction was reversed. Participants responded by selecting a number, and the next trial started automatically. It is important to emphasize that no object was retested.

The old-new assignment of objects, the viewpoint (CC or NC) chosen, and whether an image or its mirror reflection was shown, were all randomized across participants. For any participant, an "old" object's study-test image pairs were either both mirror reflected or both not reflected. In other words, no participant had to consider that an image was mirror reflected.

It took about 20 minutes for a participant to complete the experiment.

Participants

Fifty-seven fresh UCLA undergraduate students were recruited similarly as in Experiment 1.

Results

The canonical views, in the study phase, were overall rated as more attractive than non-canonical views. In the basic level block, the *mean* ratings were 0.23 and 0.09, respectively ($t(56) = 2.22$, $p = 0.03$). In the subordinate block, no difference could be found ($t < 1$).

The data in the test phase are here reported as the frequency an object was categorized as "old." For studied objects, this "old" response is the hit rate (Figure 3). For "new" objects, this "old" response is the false alarm rate (Figure 4). Our analysis of *d'* data gave rise to similar results as the analysis using hit and false alarm rates.

We categorized the rating data by positioning the decision criterion between −1 and +1, and calculated the hit and false alarm rate per participant, per block (basic vs. subordinate), per study view (canonical vs. non-canonical), and per test view (matched/same as vs. mismatched/different from the study view). First, an overall 2 × 2 × 2 ANOVA was performed on the hit data. The main effect of block was significant, $F(1, 56) = 13.86$, $p < 0.001$, where the overall hit rate for the

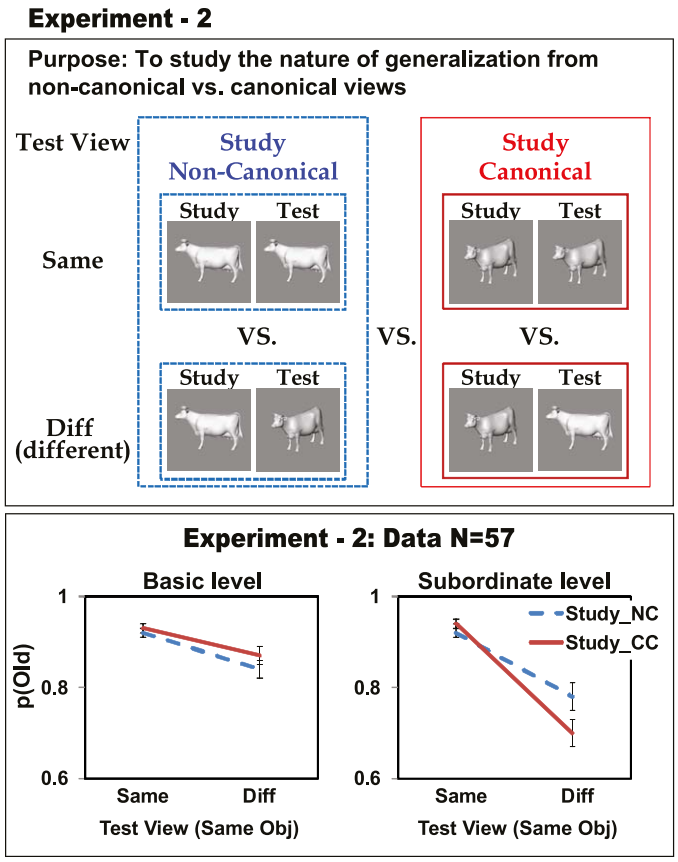

Figure 3. Experiment 2: Stimulus examples and the hit rates for the basic and subordinate blocks. In the basic level block, view generalization from matched to mismatched views was comparable, for canonical and non-canonical study views alike. In the subordinate block, interestingly, view generalization was better when non-canonical rather than canonical views were studied.

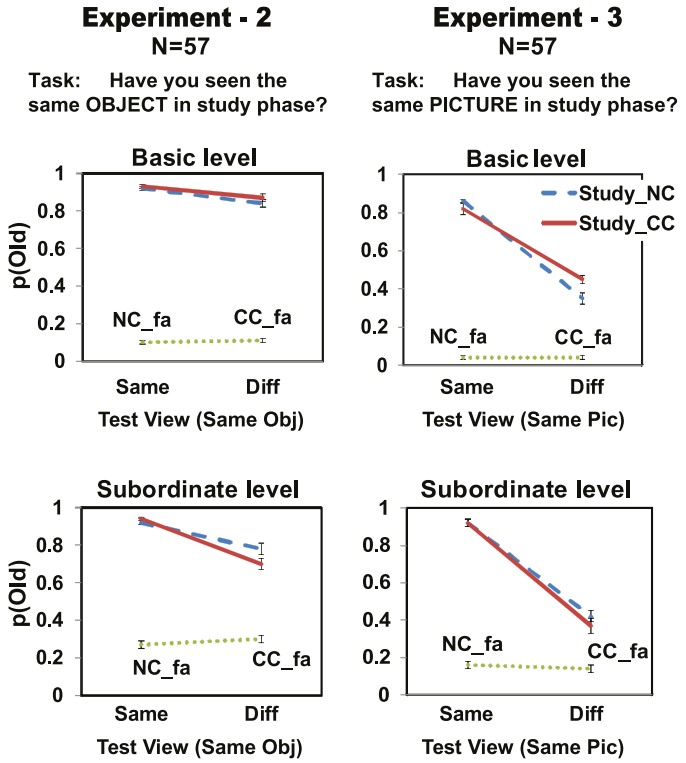

Figure 4. Comparison between viewpoint irrelevant (left, Experiment 2) and viewpoint relevant (right, Experiment 3) tasks. The green lines show false alarm rates for "new" objects. For the "old" objects, the "old" responses to a different view are lower in Experiment 3 than in Experiment 2, indicating that the participants followed experimental instructions to a certain degree. However, these "old" responses are higher than the false alarm rates for the "new" objects, implying automatic generalization from a study view to a different view. In Experiment 3 and for basic level objects, fewer false alarms were made when the objects were studied in non-canonical view and tested in a canonical view than the other way around. For subordinate level objects, the false alarms were similar for both canonical and non-canonical study views, indicating that when image-based details were made important by the task, the significant difference observed in Experiment 2 disappeared.

basic level block (89.09%) was higher than for the subordinate level (83.95%) block, not surprisingly. The main effect of test view was also significant, $F(1, 56) = 72.26$, $p < 0.001$, meaning that the overall hit rate was higher when study-test views matched than mismatched (92.97% vs. 80.07%). The main effect of study view was not significant, $F(1, 56) < 1$. There was a significant interaction between block and study view, $F(1, 56) = 5.38$, $p < 0.05$, between block and test view, $F(1, 56) = 24.09$, $p < 0.001$, and between study and test views, $F(1, 56) = 4.17$, $p < 0.05$. The three-way interaction was also significant, $F(1, 56) = 8.19$, $p < 0.01$.

In order to better understand the overall effects above, we looked at the data more closely. The hit rates were similar when the study and test views matched, for both the basic and subordinate level objects (study-view_test-view: NC_NC = 92.15%, CC_CC = 93.79%). The similar hit rates show that interestingly, the presence of within category distractors did not influence recognition when viewpoints were unchanged from study to test. No significant difference between the NC_NC and CC_CC conditions was found in either block ($t(56) = 1.16$ and 0.91, $p = 0.25$ and 0.36, for subordinate and basic levels, respectively).

These results are different from those in Gomez et al. (2008) who found that non-canonical matched views gave rise to a higher hit rate than did canonical matched views (84.80 vs. 78.90%, as compared to 92.15% vs. 93.79% in our study). The false alarm results are also different between the two studies. In Gomez et al. (2008), "new" objects tested in canonical views gave rise to a higher false alarm rate (18.20%) than in non-canonical views (10.80%). In our study, however, the main effect of "new" objects' view (canonical vs. non-canonical) was not significant in

either block (basic level: 11.26% vs. 10.54%, $t(56) < 1$; subordinate level: 30.63% vs. 26.69%, $t(56) = 1.40$, $p = 0.15$) (Figure 4). It is likely that these different results between the two studies are due to the different objects used, and the different canonical and non-canonical views selected. In a sense, the comparable performance between canonical and non-canonical matched conditions (when study-test views were identical) in our study made it easier and more interesting to interpret mismatched conditions below (when study-test views were different) for evaluating the predictions made by Equation 1. This is because any asymmetry of view generalization between canonical and non-canonical views is more likely due to the view generalization from one view to another, rather than the views themselves (Figure 3; see Discussion).

In order to test whether there was any such asymmetry, a 2 × 2 ANOVA on study (canonical, non-canonical) × test (matched, mismatched) views was performed in the basic level block. There was a main effect of test view, $F(1, 56) = 32.38$, $p < 0.001$. This means that the hit rate was higher when the study-test views matched than mismatched (NC study views: 91.84% (matched) vs. 84.41% (mismatched), CC study views: 92.93% vs. 87.18%). The main effect for study view was not significant, $F(1, 56) = 2.96$, $p = 0.09$, indicating that the hit rate was similar for CC (88.67%) and NC (88.50%) study views. The interaction was not significant, $F(1, 56) < 1$. This means that the performance drop from matched study-test views to mismatched ones was comparable regardless of view canonicality.

In comparison, the similar 2 × 2 ANOVA for the subordinate level block showed the following results. There was a main effect for test views, $F(1, 56) = 56.19$, $p < 0.001$. This again means that the hit rate was higher when the study-test views matched than when they mismatched (NC: 92.46% (matched) vs. 78.25% (mismatched), CC: 94.65% vs. 70.44%) (Figure 3). The main effect for study views was not significant, $F(1, 56) = 2.56$, $p = 0.12$, indicating that comparable hit rates were obtained regardless of the study view canonicality (NC: 85.35%, CC: 82.54%). Interestingly, there was a significant interaction between study and test views, $F(1, 56) = 7.32$, $p < 0.01$. This indicates that performance was better when a non-canonical view was studied and its canonical counterpart tested than the other way around. More specifically, the drop in hit rate from matched to mismatched views was significant for both NC and CC study views, $p < 0.001$ (CC: 24.21% drop, $t(56) = 7.65$; NC: 14.21% drop, $t(56) = 4.51$). This 10% additional drop for the CC study views indicates that, in presence of within-category objects that required additional shape details of an object to be encoded, view generalization from a non-canonical study view to a canonical test view was more accurate than the other way around.

The results from the subordinate level block of this experiment, taken in isolation, were consistent with the prediction from Equation 1. However, Equation 1 also predicts the similar pattern of results for basic level objects as well as when 3D object recognition is switched to 2D image recognition. Poggio & Edelman's model operates by matching incoming 2D images with stored 2D views. By the virtue of 2D image matching, the model attempts to find the most likely object and viewpoint that gave rise to the 2D image. Therefore, one would expect Equation 1 to make the same qualitative prediction for an image-based recognition task as explained under Scenario 2 in the Introduction. Experiment 3 tested this prediction.

Experiment 3: Viewpoint relevant recognition of 2D images

The results from the previous experiment showed that the degree of view generalization between canonical and non-canonical views was asymmetric, when detailed shape information was important for recognizing objects with within-category distractors. The goal of Experiment 3 was to check whether the asymmetric view generalization between canonical and non-canonical views held, as predicted by Poggio and Edelman (1990), when object recognition was changed to image recognition. Here, participants were instructed to remember the exact view of a study object. In the test phase, the task was to respond "old" only if the image was *exactly* the same as in the study phase (i.e., the object was being tested from the same viewpoint as it was seen in the study phase), and to respond "new" otherwise. The design of the experiment was otherwise identical to Experiment 2.

Participants and apparatus

Twenty fresh students of UCLA participated for partial course credit in undergraduate psychology courses. Thirty-seven students of the Technical University of Kaiserslautern, Germany, also participated and were paid for their time. All participants had normal or corrected-to-normal vision, were naive to the purpose of the experiment, and gave informed consent in accord with the policies of the Committee for the Protection of Human Subjects, which approved the experimental protocol at the respective universities.

The same apparatus as Experiment 1 was used at UCLA. The computer displays used in Germany were a 20″ Sun Microsystems CRT, and a 20″ Mitsubishi Diamond Pro 2070 SB CRT. The resolution was 1280 × 1024. With this resolution and image size of 450 ×

450 pixels, the angular size of the image was 13° × 13° at the viewing distance of 57 cm.

Results

The data were analyzed similarly as in Experiment 2 (Figure 4). It should be emphasized that an "old" response to an "old" object being tested in a "new" view was now a false alarm.

We first compared the overall hit rates between the basic and subordinate level blocks. The main effect of block was significant, $F(1, 56) = 4.28$, $p < 0.05$, where the overall hit rate for the basic level block was 83.91% and that for subordinate level block was 91.84%. Interestingly, the hit rate for the basic level was *lower* than that for the subordinate level block. Because the average hit rate for matched views in Experiment 2 was 92.97%, this effect in Experiment 3 must have to do with the task difference between Experiments 2 and 3. In Experiment 3, it seemed difficult for participants to remember the exact studied images in the basic level study phase without similarly shaped objects for explicit comparison. In absence of within category distractors with similar shapes and features, view generalization appears to be automatic. This automatic view generalization not only raised the false alarm rate for "old" objects' in their "new" views (see below), but also lowered the hit rates for the basic level objects.

We further looked into the decision criterion in order to check whether this lowered hit rate was due to the decision criterion change from Experiment 2 to Experiment 3. It turned out that this change, from the β value of 1.02 to 1.25, did not reach significance ($t(56) = 1.83$, $p = 0.073$). In terms of the bias, however, the bias difference between the two experiments was highly significant. This difference, however, was almost entirely due to the shift of the bias-free or optimal decision criterion location from Experiment 2 to Experiment 3. This is because in Experiment 2 the relative frequency of "signal" and "noise" was 1:1, whereas in Experiment 3 it was 1:3. As a result, the optimal decision criterion was shifted to the right from Experiment 2 to Experiment 3. Nevertheless, this bias difference should not affect any of the conclusions in this study, because the hypothesis testing was based on the pattern of results within each experiment.

We also calculated d' per participant, per block, and per study view (canonical or non-canonical). It turned out that in the basic level block, the average d' in Experiment 3 was significantly lower than in Experiment 2 (2.12 vs. 2.79, $p < 0.0001$). However, the d' scores in the subordinate block were comparable between the two experiments (1.83 vs. 1.70, $p = 0.14$). These results suggest that, from Experiment 2 to Experiment 3, participants could not possibly just shift the decision criterion in order to switch from object recognition to image recognition. Instead, the way in which internal representations were constructed was different between the two experiments. In other words, the independence model proposed by Poggio and Edelman (1990) was unlikely to account for these results.

Within each block, the hit rates were similar for non-canonical and canonical matched views (basic level: 85.96% and 81.86%, $t(56) = 1.89$, $p = 0.63$; subordinate level: 91.67% and 92.02%, $t(56) < 1$). The overall false alarm rate for "new" objects was lower for the basic level (4.01%) than for the subordinate level block (15.08%), an effect similar to what was observed for the object-based task in Experiment 2. Within a block, there was no significant difference in the overall false alarm rates for new objects between canonical and non-canonical views (basic level, 3.80% and 4.23%, $t(56) < 1$; subordinate-level, 13.36% and 16.81%, $t(56) = 1.47$, $p = 0.15$).

We further separated the false alarm rates for the "old" and "new" objects. There was a significant difference between these two sets of objects for both non-canonical and canonical test views in each block. In the basic level block, the false alarms for the non-canonical test views were 35.46% ("old" objects) vs. 4.23% ("new" objects), $t(56) = 11.86$, $p < 0.001$. The false alarms for the canonical test views were 45.15% ("old" objects) vs. 3.80% ("new" objects), $t(56) = 16.68$; $p < 0.001$. In the subordinate level block, the false alarms for the non-canonical views were 42.46% ("old" objects) vs. 16.81% ("new" objects), $t(56) = 7.32$, $p < 0.001$. The false alarms for the canonical views were 36.93% ("old" objects) vs. 13.36% ("new" objects), $t(56) = 7.20$, $p < 0.001$. These differences further indicate that there was an automatic view generalization from a study view to other views, even though the participants were explicitly instructed not to generalize. For "old" objects, the false alarms for non-canonical views (42.26%) were not significantly different from canonical views (36.93%), $t(56) = 1.24$, $p = 0.22$. However, the "old" responses for "old" objects were always lower when the study-test views were mismatched rather than matched. We refer to this difference as the performance drop. The average drop was bigger in Experiment 3, from 87.88% to 40.00%, $F(1, 56) = 72.26$, $p < 0.001$ than in Experiment 2 (from 92.97% to 80.07%), indicating that the participants followed the instructions to perform the image recognition task to some extent, despite the automatic view generalization (Figure 4).

We looked into the performance drop more closely. The main effect of study view was not significant, $F(1, 56) < 1$. There was a significant interaction between block and study view, $F(1, 56) = 5.14$, $p < 0.05$, between block and test view, $F(1, 56) = 7.18$, $p < 0.01$, but not between study and test views, $F(1, 56) = 1.63$, $p = 0.20$.

The three way interaction between block, study view, and test view was also significant, $F(1, 56) = 8.00$, $p < 0.01$.

In order to better understand the above interaction effects, we analyzed the "old" response data within each block. A 2 × 2 ANOVA for the basic level block showed that there was a main effect for test view, $F(1, 56) = 270.78$, $p < 0.001$. This means that the "old" response rate was higher when the study-test views matched than when mismatched (NC: 85.96% vs. 35.46%, CC: 81.86% vs. 45.15%). The main effect for study view was significant, $F(1, 56) = 4.47$, $p < 0.05$. This main effect is complex to interpret, because the comparison involves both hits and false alarms. The interaction between study and test views was highly significant, $F(1, 56) = 14.41$, $p < 0.001$. This means that the performance dropped from matched to mismatched views significantly differently (50.49% drop for non-canonical study views, and 36.71% for canonical ones). The results indicate that participants made fewer errors in following the instructions for non-canonical objects, implying that either non-canonical images were encoded in memory better or canonical images generalized to other views more involuntarily. This involuntary view generalization, in absence of within category distractors, probably led to more false alarms for "new" views of "old" objects.

The similar 2 × 2 ANOVA for the subordinate level block, in comparison, showed the following results. There was a main effect for test view, $F(1, 56) = 385.93$, $p < 0.001$, as expected. There was no main effect for study view, $F(1, 56) = 1.31$, $p = 0.25$. There was no significant interaction between study and test views, $F(1, 56) = 1.14$, $p = 0.29$. This means that the performance drop was comparable from a non-canonical study view to a canonical test view (49.21%) and vice versa (55.09%).

The results above indicated that the effects found in Experiment 2 were reversed when the task of viewpoint irrelevant object recognition was changed to view dependent 2D image recognition of 3D objects.

Finally, the attractiveness rating for this experiment showed a pattern similar to that in Experiment 2. In the basic level block, canonical views were again rated as more attractive than non-canonical views, 0.61 vs. 0.36, $t(56) = 3.90$, $p < 0.001$. In the subordinate block, no difference could be found ($t < 1$). Thus the attractiveness rating cannot explain the difference in results found in the two experiments.

In conclusion, results in Experiment 3 were inconsistent with Equation 1, which never predicts better view generalization from canonical to non-canonical views than the other way around. The fact that a "new" view of an "old" object was involuntarily categorized as "old" substantially more often than chance also argues against the independence model by Poggio and Edelman (1990).

Discussion

Successful object recognition requires an efficient internal representation and matching process between the representation and the input. Ideally, the representation should accurately characterize the full 3D structure of an object to allow recognition from any viewpoint. In practice, however, the visual system may not be powerful enough to reconstruct the full 3D structure from a finite number of views. Thus, this question remains: what can the visual system do to approximate the 3D structure of an object so that the object from a new viewpoint can be effectively inferred, albeit imperfectly?

Poggio and Edelman (1990) proposed their model to achieve viewpoint invariant recognition by summing similarity measures between the input image and each independently stored view of the object (Longuet-Higgins, 1990). When viewpoint relevant recognition is required, it is sensible not to sum but to select the best possible match from the stored views to compare with the threshold criterion. In both view relevant and irrelevant object recognition, Poggio and Edelman (1990) would predict that Equation 1 hold, i.e., recognition performance should be better for a canonical view that follows a non-canonical view presentation than the other way around. A subset of our experimental results, however, was inconsistent with this prediction even with our assumption of a generic similarity measure in the Poggio and Edelman (1990) model. This suggests that the assumption of independence of stored views made by this model cannot account for human performance.

Alternatives to view-approximation model

The inadequacy of the independently stored views assumption has also been experimentally demonstrated by other studies. Wallis and Bülthoff (2001) have shown that observers tend to better associate sequentially presented views with a single object when the sequence in time was smooth rather than scrambled. This result suggests that coherent and smooth transition from one view to its neighboring view is relevant and important in building a representation. In other words, independent storage of views is insufficient; and encoding the transition from one view to the next, or effectively encoding object rotation in 3D from one view to the next, may be necessary, even if this encoding is imprecise.

If the internal representation of an object is not encoded as an independent set of views, then an incoming view needs to be integrated into the already existing representation of that object. How might an existing representation incorporate a new input view in order to infer the appearance of the object from a viewpoint that is rotated away from the new input? Unfortunately, no computational models exist that offer a detailed account on this integration. We can only speculate how this task might be achieved from existing models.

Alter and Jacobs (1998) have studied a simpler problem, when there is a 3D object model and a 2D input image, both specified by coordinates of point features. They asked the following question: how can correspondence between the 3D model and the 2D image be fully established, under the assumption that correspondence between some feature points is already known? Alter and Jacobs (1998) characterized this problem as uncertainty propagation from the known correspondence to the unknown. Uncertainty exists because the exact pose of the 3D object from which the 2D image is projected is unknown, and because there are always measurement errors in locating the image features. Undoubtedly, the problem becomes much more challenging when some correspondence is known from one viewpoint, and new correspondence needs to be established from a different viewpoint, when the 3D structure of the object is not perfectly represented.

Another alternative to the view-approximation model are the feature-based models, which are promising in explaining biological object recognition (Ullman, 2006; Torralba, Murphy, & Freeman, 2007; Wallis, Siebeck, Swann, Blanz, & Bülthoff, 2008). It is likely that a canonical view may share more features with other views of the same object than a non-canonical view does. Thus, feature-based models may be able to offer an explanation to the asymmetric results between canonical and non-canonical views found in the current study. Whether or not existing feature-based models can indeed predict the patterns of the results in the current study, however, can only be known after model simulations, which is beyond the scope of this study.

Involuntary view generalization

In Experiment 3, object recognition is viewpoint relevant. Therefore, when an object is seen from a certain viewpoint, view propagation is supposed to be minimal. The existing representation is also supposed to be subdued since the major source of information for the viewpoint relevant recognition is primarily the study view. Our results indicated that the effects found in Experiment 2 were reversed when the task of viewpoint irrelevant object recognition was changed to view dependent 2D image recognition of 3D objects. We hypothesize that in the natural task in Experiment 2 information from one previously seen view of an object automatically propagates to nearby viewpoints of the same object. That is, the representation of the object automatically attempts to predict, or characterize, what the object looks like from other viewpoints. As far as we know, this is the first time that unavoidable interference from object-judgment to image-judgment is being explicitly reported and discussed in object recognition literature.

We found in Experiment 2 that for basic level objects, this propagation was symmetric between canonical and non-canonical views. Importantly, we also found that for subordinate level objects, the propagation from non-canonical to canonical views was more accurate than the other way around. In Experiment 3, when participants were instructed to treat a new view of an "old" object as "new" instead of "old," they were effectively instructed to suppress the process of view propagation that probably led to reconstruction of the internal representation. Apparently, such suppression was not fully effective. Some aspects of this view propagation were involuntary in that "new" views of "old" objects' were categorized as "old" more often than "new" objects. Based on our results it appears that the involuntary view propagation is more likely for objects studied in canonical than in non-canonical views, especially so in absence of within category distractors. On the other hand, the experimental instruction, or the top-down aspect of the suppression, was partially effective in that for the subordinate level objects, the asymmetric effect of view propagation disappeared. In other words, since the propagation from non-canonical to canonical views was more accurate in Experiment 2, it means that the suppression in Experiment 3 was relatively more effective for non-canonical views. The results, similar to those for the basic level in Experiment 3, indicate that participants made fewer errors in following the instructions for non-canonical views, probably implying that non-canonical images were better encoded into memory. Performance was better for images studied in non-canonical views both in Experiment 2 (object recognition task) and Experiment 3 (image recognition task). On the other hand, the same results may also imply that the involuntary view generalization from canonical to other views can be suppressed when image details are made more important for the task at hand, thereby reducing the asymmetry of results seen in the subordinate level block for Experiment 3. But the reason for this differential suppression remains unclear based on the results of the experiments being reported here. Nevertheless, it is reasonable to assume that such suppression works similarly for both subordinate and

basic level objects. Then one would expect that, for basic level objects, propagation from non-canonical views was also more effectively suppressed. This is consistent with the data in Experiment 3.

In summary, we studied internal shape representations in object recognition with the following manipulations. We manipulated how informative an object appeared by presenting the object either from a canonical or a non-canonical view. We manipulated the task demand by asking a participant to recognize an object with either between-category or within-category distractors. We also manipulated the task demand by asking a participant to recognize an object either regardless the viewpoint that had been seen or from the exact viewpoint that had been seen. Our results showed that the view-approximation model could not fully account for our experimental data, and suggest that rotational relationship from one view to another needs to be incorporated in the internal representation.

Acknowledgments

This research was in part supported by a Marie Curie Career Integration Grant (#293901) from the European Union awarded to Tandra Ghose and a US NSF grant (BCS 0617628) to Zili Liu. Part of this research was presented at the European Conference on Visual Perception, Lausanne, Switzerland, 2010; and at the Vision Sciences Society, Naples, Florida, 2011. We thank Drs. David Bennett and Nestor Matthews for helpful comments and suggestions, who read the entire manuscript. We also thank the anonymous reviewers and, in particular, Shimon Ullman, the editor, for his insightful comments and suggestions.

Commercial relationships: none.
Corresponding author: Tandra Ghose.
Email: tandra@berkeley.edu.
Address: Department of Psychology, Technical University of Kaiserslautern, Germany.

References

Alter, T. D., & Jacobs, D. W. (1998). Uncertainty propagation in model-based recognition. *International Journal of Computer Vision, 27*, 127–159.

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

Biederman, I. (2000). Recognizing depth-rotated objects: a review of recent research and theory. *Spatial Vision, 13*(2-3):241–253.

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

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

Binford, T. O. (1971). Visual perception by computer. *IEEE Conference on Systems Science and Cybernetics,* Miami, FL.

Blanz, V., Tarr, M. J., Bülthoff, H. H., & Vetter, T. (1999). What object attributes determine canonical views? *Perception, 28*, 575–599.

Brainard, D. J. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science of the USA, 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.

Cutzu, F., & Edelman, S. (1994). Canonical views in object representation and recognition. *Vision Research, 34*, 3037–3056.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences, 21*(4): 449–467.

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research, 32*, 2385–2400.

Edelman, S., & Poggio, T. (1992). Bringing the grandmother back into the picture: a memory-based view of object recognition. *International Journal of Pattern Recognition & Artificial Intelligence, 6*, 37–62.

Edelman, S., & Shahbazi, R. (2012). Renewing the respect for similarity. *Frontiers in Computational Neuroscience, 6*(45), 1–19, doi: 10.3389/fncom.2012.00045.

Gomez, P., Shutter, J., & Rouder, J. N. (2008). Memory for objects in canonical and noncanonical viewpoints. *Psychonomic Bulletin & Review, 15*, 940–944.

Grenander, U. (1993). *General Pattern Theory*. New York: Oxford University Press.

Harries, M. H., Perrett, D. I., & Lavender, A. (1991).

Preferential inspection of views of 3-D model heads. *Perception, 20*, 669–680.

Hayward, W. G. (2003). After the viewpoint debate: where next in object recognition? *Trends in Cognitive Sciences, 7*(10):425–427.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271–304.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology, 13*(2):1–9.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.

Koenderink, J. J., & van Doorn, A. J. (1976). The singularities of the visual mapping. *Biological Cybernetics, 24*(1):51–59.

Liu, Z. (1996). Viewpoint-dependency in object representation and recognition. *Spatial Vision*, (Special Issue on Perceptual Learning and Adaptation in Man and Machine), *9*, 491–521.

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object Classification for Human and Ideal Observers. *Vision Research, 35*, 549–568.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience, 19*, 577–621.

Longuet-Higgins, H. C. (1990). Recognizing three dimensions. *Nature, 343*, 214–215.

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

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

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

Newell, F. N., Ernst, M. O., Tjan, B. S., & Bülthoff. H. H. (2001). Viewpoint dependence in visual and haptic object recognition. *Psychological Science, 12*(1):37–42.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance*, Chapter 9 (pp. 135–151). Hillsdale, NJ: Lawrence Erlbaum.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience, 5*(4): 291–303.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10,* 437–442.

Perrett, D. I., & Harries, M. H. (1988). Characteristic views and the visual inspection of simple faceted and smooth objects: "tetrahedra and potatoes." *Perception, 17*, 703–720.

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*, 497–515.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology, LV*, 899–910. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

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

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

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3*, 1199 –1204.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 21*, 233–282.

Tarr, M. J., & Bülthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1494–1505.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey, and machine. *Cognition, Special Issue on Image-based Object Recognition, 67*(1/2):1–20.

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

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience, 1*(4): 275–277.

Tarr object database. (http://www.tarrlab.org/). Pittsburgh, PA: Carnegie Mellon University.

Tjan, B. S. & Legge, G.E. (1998). . The viewpoint complexity of an object-recognition task. *Vision Research, 38*(15-16):2335–2350.

Torralba, A., Murphy, K. P., & Freeman, W. T. (2007). Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*, 854–869.

Ullman, S. (1996). *High-level vision: Object recognition and visual cognition*. Cambridge, MA: MIT Press.

Ullman, S. (2006). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences, 11*, 58–64.

Ullman, S., & Basri, R. (1989). Recognition by linear combinations of models. Massachusetts Institute of Technology. AI Memo 1052.

Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*, 992–1006.

Verfaillie, K., & Boutsen, L. (1995). A corpus of 714 full-color images of depth-rotated objects. *Perception & Psychophysics, 57*, 925–961.

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences of the USA, 98*, 4800–4804.

Wallis, G., Siebeck, U., Swann, K., Blanz, V., & Bülthoff, H. H. (2008). The prototype effect revisited: Evidence for an abstract feature model of face recognition. *Journal of Vision, 8*(3):20, 1–15. http://www.journalofvision.org/content/8/3/20, doi:10.1167/8.3.20. [PubMed] [Article]

Warrington, E. K., & Taylor, A. M. (1973). The contribution of the right parietal lobe to object recognition. *Cortex, 9*(2):152–164.

Weinshall, D., & Werman, M. (1997). On view likelihood and stability. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*, 97–108.