PERGAMON

Vision Research 38 (1998) 2259–2275

Vision Research

# Why the visual recognition system might encode the effects of illumination

Michael J. Tarr [a,*], Daniel Kersten [b], Heinrich H. Bülthoff [c]

[a] *Department of Cognitive and Linguistic Sciences, Brown University, Box 1978, Providence, RI 02912, USA*
[b] *Department of Psychology, University of Minnesota, N218 Elliot Hall, 75 East River Road, Minneapolis, MN 55455, USA*
[c] *Max-Planck Institute for Biological Cybernetics, Tübingen, Germany*

Received 9 December 1997; received in revised form 12 January 1998

**Abstract**

A key problem in recognition is that the image of an object depends on the lighting conditions. We investigated whether recognition is sensitive to illumination using 3-D objects that were lit from either the left or right, varying both the shading and the cast shadows. In experiments 1 and 2 participants judged whether two sequentially presented objects were the same regardless of illumination. Experiment 1 used six objects that were easily discriminated and that were rendered with cast shadows. While no cost was found in sensitivity, there was a response time cost over a change in lighting direction. Experiment 2 included six additional objects that were similar to the original six objects making recognition more difficult. The objects were rendered with cast shadows, no shadows, and as a control, white shadows. With normal shadows a change in lighting direction produced costs in both sensitivity and response times. With white shadows there was a much larger cost in sensitivity and a comparable cost in response times. Without cast shadows there was no cost in either measure, but the overall performance was poorer. Experiment 3 used a naming task in which names were assigned to six objects rendered with cast shadows. Participants practised identifying the objects in two viewpoints lit from a single lighting direction. Viewpoint and illumination invariance were then tested over new viewpoints and illuminations. Costs in both sensitivity and response time were found for naming the familiar objects in unfamiliar lighting directions regardless of whether the viewpoint was familiar or unfamiliar. Together these results suggest that illumination effects such as shadow edges: (1) affect visual memory; (2) serve the function of making unambigous the three-dimensional shape; and (3) are modeled with respect to object shape, rather than simply encoded in terms of their effects in the image. © 1998 Elsevier Science Ltd. All rights reserved.

*Keywords:* 3-D objects; Illumination; Visual recognition system

## 1. Why the visual recognition system might encode the effects of illumination

How can objects be recognized given that images differ drastically depending on the illumination? The standard answer is that, early on, the visual system extracts illumination-invariant features, such as edges and contours, but discounts 'spurious' features, such as shadows and specularities [1,2]. Recent 'image-based' theories of object recognition [3–6], however, propose that object representations may be tied more closely to the original image. This is particularly the case given that illumination features such as cast shadows are difficult to discount using low-level mechanisms [7]. Thus, visual recognition may be sensitive to the illumination conditions (or it's consequences) under which an object is learned.

The variation in the image due to a change in illumination direction appears small compared to the variation that arises from a change in viewpoint or configuration. However, an examination of the consequences of varying the direction of illumination reveals that this is not the case (Fig. 1)—there are many instances where varying illumination direction produces a larger change in the image than does varying viewpoint [8].

The way in which lighting affects the image of an object is extraordinarily complex involving changes in overall magnitude, shading, and shadows. Whether or

* Corresponding author. Tel.: +1 401 8631148; fax: +1 401 8632255; e-mail: michael_tarr@brown.edu.

0042-6989/98/$19.00 © 1998 Elsevier Science Ltd. All rights reserved.
PII: S0042-6989(98)00041-8

not the visual system discounts each of these effects depends on the value of the information for recognition or other visual tasks. Let us consider three distinct examples. First, the mean level of luminance on the image varies with the overall magnitude of illumination. Because it is generally assumed that shape is a primary determinant in object perception, it is not surprising that human vision discounts slow variations in illumination magnitude at the retina. Second, shading (variations in intensity inside the object contours) is difficult to discount, yet is potentially useful because it results from the local interaction between the surface orientation and light source direction, consequently, shading may provide information about shape. Indeed, there is a long history of recovering three-dimensional shape-from-shading [9,10]. Thus it would seem that human vision should represent shading at some level [11]. Third, cast shadows result from the interaction between a light source, the casting object, and the receiving surface. Unlike shading, the form of a cast shadow is not locally determined and is affected by the spatial characteristics of surfaces distant to the object of interest (in this paper, we will consider only intrinsic shadows—those cast on an object by itself). Thus, cast shadows are perhaps even more difficult to discount than shading, but again provide potentially useful information about the shape, and, in particular, the three-dimensional structure [12]. Support for this idea in recent work has shown that both attached [13,14] and cast shadows [15] provide some information regarding the three-dimensional shape.

Given that there is useful information in both shading and cast shadows, it seems reasonable to consider approaches to object representation that do not discount these effects at an early level. However, images of objects are highly illumination specific, therefore, once we have introduced shading and cast shadows into the representation, they may be difficult to account for without the use of higher-level and top-down mechanisms [16]. If this is the case, then variation in illumination, much as with the variation in viewpoint [17–20], may have consequences for the speed and accuracy of recognition. Specifically, while variation in illumination may hinder recognition, preserving the effects of illumination may sometimes facilitate recognition. For example, early-level filtering such as edge detection can increase the similarities between object representations (because luminance edges may arise from different causes, e.g. albedo, highlights, or shape, increasing the probability of false correspondences). Given that we generally expect recognition to become more difficult as object representations become increasingly similar [21,22], then the early discounting of the effects of illumination may actually lead to poorer recognition. Thus, there is an inherent trade-off between the information that may be useful for discriminating between objects, and the ambiguities inherent in interpreting encoded information. Consider the images in Fig. 2, and the following two cases for encoding the image information.

### 1.1. Edge-based

In one case, we could imagine a ‘smart’ edge-detector that represents image information about an object as a line drawing that marks only significant surface edges—for the objects in Fig. 2, occlusion and orientation discontinuities. Shading gradients, attached shadows, and cast shadow boundaries are filtered out. This representation would be completely robust over illumination variability with regard to recognition. Of course, a perfect edge-detector does not currently exist, therefore any real-world edge-based model would be expected to show some illumination sensitivity. On the other hand, there are edge-based models of human object recognition that explicitly rely on precise line

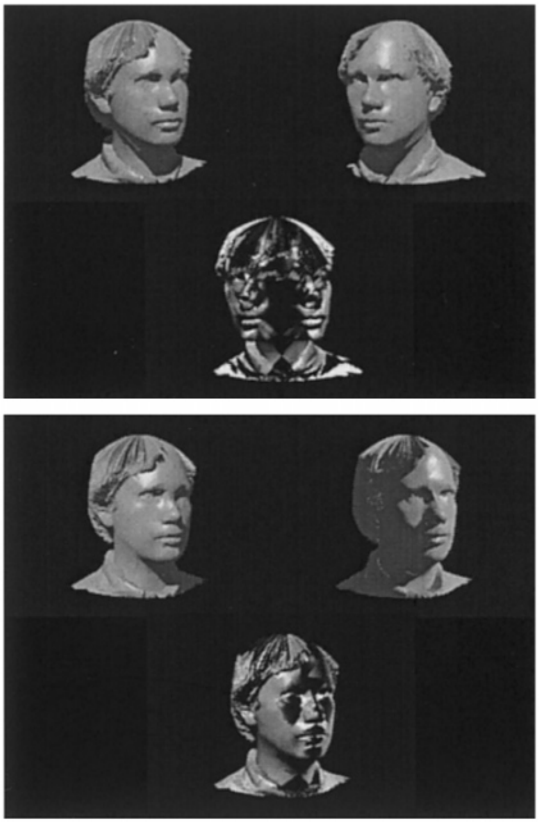

Fig. 1. The distance, in pixel intensity space, between the pictures of two viewpoints of the same head in the same illumination (top panel) is actually less than the distance between the pictures of two illumination directions of the same head in the same viewpoint (bottom panel). In the bottom image of each panel, the pixel brightness is proportional to the squared difference between the left and right images. The distance and between the two vectors representing the two views of the same face for a fixed illumination is 62 (arbitrary units) and 38°; the distance and angle between the two images with a change in illumination direction for a fixed view is 65 and 42°.

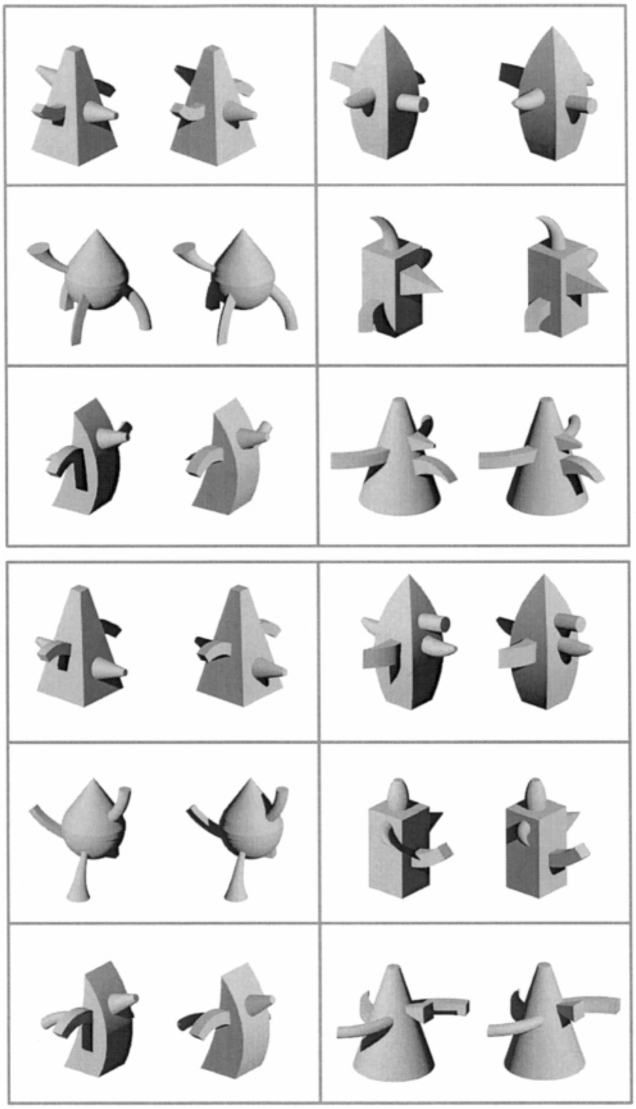

Fig. 2. The sets of novel three-dimensional objects used in experiments 1–3. The top set of 6 objects was used in experiment 1. Objects were illuminated with ambient and point light sources from either the left or the right (images from both illumination directions are shown for each object). The objects in the top set each contain a qualitatively unique central volume and an unique configuration of attached parts. The top and bottom sets of objects were used together in experiments 2 and 3. Because no object in this combined set contains a qualitatively unique central volume, the objects could be discriminated only by attending to the configuration of the parts. The objects were created by Scott Yu.

drawings as input (in part in order to achieve invariance over illumination [1,23,24]). Such representations would also be impoverished because illumination features, such as shading and shadows, are no longer available for distinguishing subtle three-dimensional variations between the objects. Recognition would have to rely entirely on the geometrical distinctions remaining after projection. As a result, accurate decisions could require more time because of the greater similarity between representations in memory.

### *1.2. Image-based*

In the second case, we could imagine that object contours, as well as informative shading (i.e. contrast gradients), and cast shadow markings are represented in the visual memory. Such a representation would be sensitive to illumination change. However, representations encoding the effects of illumination would have greater potential for distinguishing subtle three-dimensional variations between objects in the absence of other geometrical features. Thus, we would expect an increase in overall recognition performance.

Although the above cases certainly do not represent all of the possible treatments of the effects of illumination, they do present two plausible and typical examples of the kinds of models that have been proposed for object recognition [25,3]. We will therefore use them as anchor points for our investigation of the effects of illumination in object recognition. Using computer graphics psychophysics to synthesize novel three-dimensional objects we can manipulate a variety of factors that may help to distinguish between the two cases outlined above. The primary manipulation used within all of the experiments is to have observers study objects illuminated from one direction and then test observers on the same objects illuminated from a different illumination direction. This simple manipulation allows us to assess whether or not the object representations are illumination specific, that is, the degree to which the original illumination conditions are encoded in the visual memory. In Experiments 1 and 2 this illumination variation will occur over two sequentially presented images, and hence, will explore only short-term visual memory. In Experiment 3 illumination variation will occur over objects learned on one day and tested on a second day, and hence will explore long-term visual memory. A second primary manipulation is used between Experiments 1 and 2: we varied the objective similarity between objects in two ways. First, we increased the geometrical similarity between the objects by introducing subtle changes in the configurations of the parts. Second, we increased the photometric similarity by removing some of the effects of illumination, in particular, whether cast shadows were present or not. Finally, in Experiment 3 we explored the interaction between viewpoint change and illumination change. Of particular interest was whether illumination-specific effects were also specific to familiar viewpoints. If illumination sensitivity was found to be viewpoint specific, then this would suggest that what is encoded are the effects of illumination in the image. In contrast, if illumination sensitivity was found to generalize to new viewpoints, then this would suggest that the effects of illumination are modeled in some fashion, either explicitly (e.g. in terms of illumination direction) or implicitly (e.g. as sets of image basis functions over illumination variation [26–30].

## 2. Experiment 1

To investigate the role of illumination in object recognition, we began with a straightforward same/different discrimination between pairs of sequentially presented images. This sequential-matching paradigm has been quite popular in the object recognition literature and has been used in numerous studies investigating sensitivity to viewpoint [31,32], as well as other properties of objects [33]. The advantage of this paradigm is that the experimenter can easily manipulate the information available between study and test, thereby assessing how recognition is affected by variation in the image. Moreover, because each image pair represents an unique encoding condition and memory test, it is possible to repeat objects and conditions to gain statistical power. On the other hand, the sequential-matching paradigm does leave open the question as to whether we are measuring visual memory, albeit short-term, or rather, are tapping into more transient 'sensory' representations that do not reflect the kinds of information encoded at any level in the visual memory. Arguing against this possibility, there are many instances where experimenters have used similar experimental parameters in a sequential-matching task and have obtained relatively invariant performance. Therefore, given adequate pictorial masks for each image and sufficient time between image presentations (as measured by Ellis and Allport [33]) one can be reasonably sure that we are assessing visual memory and not sensory processing that is task-specific[1].

To assess the impact of illumination variation on object recognition, we used the six objects shown in the top panel of Fig. 2 in a sequential-matching paradigm (illustrated in Fig. 3). In each trial an image of an object was briefly presented, followed by a random parts mask (illuminated from both directions), and then a second object, also masked. The crucial manipulation was whether there was a change in illumination across the two objects (as shown in the left and right images for each object in Fig. 2). Observers judged whether the objects were the same or different regardless of illumination. Note that the use of these six objects provides a relatively strong test of whether the effects of illumination are encoded in the visual memory. Given the extreme qualitative shape differences between each of the objects in the set, there is little reason under the edge-based case to expect observers to use anything other than shape-based recognition mechanisms. Indeed, it is generally thought that recognition performance becomes more invariant over transformations of the image as the similarity between objects decreases [21,22]. Thus, if we find a reliable cost for changing the illumination between study and test, this would provide some support for the image-based case.

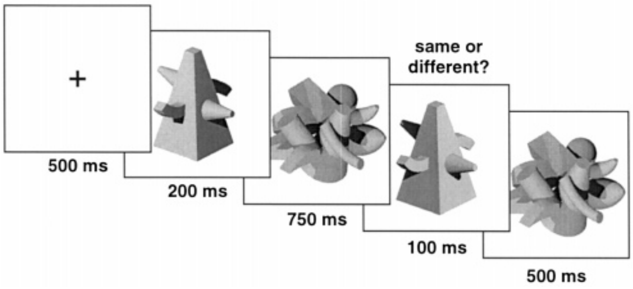

Fig. 3. The sequential-matching paradigm used in experiments 1 and 2. Each trial was composed of a fixation cross, an image of an object, a mask, a second image of an object, and a mask. Presentation times are as shown [33]. The mask was created by randomly intermixing all of the individual parts present in the stimulus objects and illuminating them from both lighting directions. The mask used in the NoShadow condition of experiment 2 was rendered without cast shadows.

### 2.1. Method

#### 2.1.1. Participants

Thirty-two Yale undergraduates participated in return for course credit or pay. None of the participants had seen the stimuli prior to the experiment.

#### 2.1.2. Materials

Six novel objects were synthesized by compositing simple three-dimensional volumes and by illuminating them with ambient and point light sources (Fig. 2, top panel). Several properties of these objects were crucial. First, the objects were composed of parts so as to produce intrinsic shadows. Second, the objects were novel to reduce the effect of top-down mechanisms in resolving shadow identity [16]. Third, the objects were rendered with uniform albedo to reduce the effect of albedo/shadow ambiguity. Fourth, the objects were illuminated from either the left or the right, avoiding the effects of up/down illumination changes on perceived shape[2].

[1] One related concern is that even given adequate masking, it is conceivable that, as claimed by Biederman and Cooper [34], observers are relying on local differences between the images, in particular, due to an interaction between the ventral and dorsal visual pathways. At least two points argue against this account. First, one possible method for addressing this issue is to include an image-plane translation between image pairs. However, given the strong evidence forcomplete translation invariance in similar tasks [35], it seems rather unlikely that this manipulation would alter behavior with regard to other image variation. Furthermore, in a sequential-matching experiment using faces that were rotated or scaled between study and test, Braje et al. [36] found costs for illumination variation even larger than those reported here. Second, there is no strong evidence that the action channel plays any role in object recognition, and, indeed, there is some evidence that it does not, being a purely 'on-line' system [37].

[2] As demonstrated by Brewster [38] and Ramachandran [39] changes in illumination vertically give rise to the well-known convex/concave switch for a shaded hemisphere, in part due to the strong priority on lighting from above. Likewise, Johnson et al. [40] demonstrated that lighting from below can be deleterious to the recognition of familiar objects such as faces.

In terms of shape, each of the objects contained both a qualitatively unique central volume and a qualitatively unique configuration of attached parts. Thus, the shapes of the objects were relatively dissimilar. The objects were illuminated from either the left or from the right, varying both the shading and the shadows cast on the objects. Lighting conditions were created as follows. The objects rested on an $x$–$z$ plane, with $z$ pointing out of the picture; $x$ and $y$ were the horizontal and vertical directions, respectively. The camera's viewpoint was oriented 25° about the $x$-axis, above the $x$–$z$ plane. The light source was a point 40° about the $x$-axis above the $x$–$z$ plane. It was rotated either $+30°$ or $-30°$ about the $y$-axis corresponding to right and left illumination, respectively. Objects were rendered with ray-tracing, producing sharp cast shadows. The reflectance model used an ambient reflectance of 0.15, and a Lambertian reflectance of 0.6; there was no specular term. The objects were rendered under orthographic projection using Wavefront's 'The Advanced Visualizer 4.1' (Santa Barbara, CA) running on a Silicon Graphics workstation. The images were then transferred to an Apple Macintosh and converted to 8-bit greyscale at 72 dpi for presentation.

### 2.2. Design and procedure

A sequential-matching paradigm was used in Experiment 1. Participants had to judge whether two sequentially presented objects were the same or different. Each participant ran 240 test trials in which the 6 objects from the top panel of Fig. 2 were shown in 2 illumination directions as described above. The objects appeared equally often in each illumination condition (20 times each). Each trial was composed of a fixation cross for 500 ms, an image of an object for 200 ms, a visual mask for 750 ms, a second image of an object for 100 ms, and the same mask for 500 ms (Fig. 3). These times were chosen so as to minimize the likelihood that participants would rely on local differences in image properties, but, rather, would encode the first image in the short-term visual memory for subsequent matching to the second image [33]. The participants' task was to judge as quickly and as accurately as possible whether the two objects in a given trial were the same or different regardless of any change in illumination. From the time the second object image was presented participants were given 1600 ms in which to respond. If they did not respond by this deadline or they responded incorrectly, they heard a loud beep as feedback[3]. One half of the trials paired an object with itself (same response) in one of two illuminations (appearing equally often) and one half of the trials paired an object with a different object (different response) in one of two illuminations. Similarly, one half of the trials paired an image showing one illumination direction with an image showing the same illumination direction (NoChange condition) and one half of the trials paired an image showing one illumination direction with an image showing the other illumination direction (Change condition). Participants responded by pressing one of two keys on a keyboard. Responses and response times were recorded using custom-designed presentation software running on a Macintosh LC475. The Macintosh was also used to control the stimulus presentation at a resolution of 72 dpi on an Apple 13 in. color monitor. Participants viewed the objects binocularly from a distance of approximately 60 cm from the screen resulting in images (which were not presented in stereo) that subtended a region of approximately $5.7° \times 5.7°$ of visual angle. Images were presented in synchronization with the refresh of the screen and were preloaded into the computer memory so that the entire image appeared in one refresh cycle. The order of presentation of the trials was randomized for each participant and participants received two rests at random intervals. The entire experiment took less than 1 h.

### 2.3. Results and discussion

For the purposes of computing mean response times, incorrect responses were discarded. No adjustments were made to correct for outliers in that the response times were normally distributed. Mean response times were computed for same and different trials for the NoChange and Change in illumination conditions (same/different): NoChange, 807 ms/846 ms; Change, 828 ms/839 ms. As shown in Fig. 4, a change in the direction of illumination produced a reliable 21 ms cost in judging that two images are the same object, $F(1, 31) = 8.67$, $P < 0.01$. The similarity in same and different response times also allowed us to compute a sensitivity measure, *d'*, from the correct responses for same trials (hits) and the incorrect responses for different trials (false alarms) for each illumination condition: NoChange, 3.55; Change, 3.78. This is shown in Fig. 4, where a change in the direction of illumination did not produce a reliable difference in the ability of observers to judge that two images were the same object, $F(1, 31) = 1.51$, ns.

Two results stand out in Experiment 1. First, as expected based on the low shape similarity between the

[3] Feedback did not result in the participants learning the illumination directions. For this and subsequent experiments using the sequential-matching paradigm, comparisons between the first and second halves of the test session revealed no interaction between illumination conditions and position within the session.

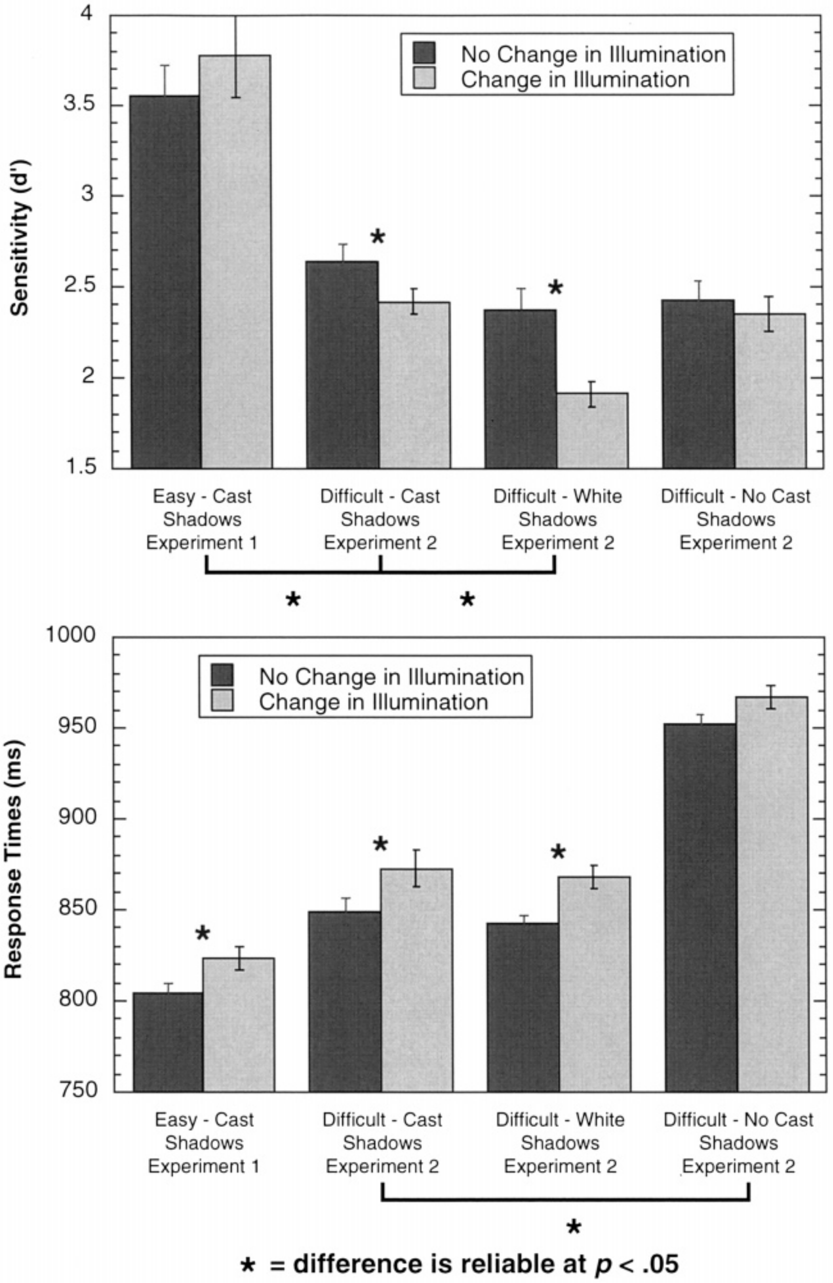

Fig. 4. Experiments 1–2. Mean sensitivity (*d'*) and mean response times for correct responses for same trials across the NoChange and Change in illumination conditions. From left to right there is a progression in the degree of perceptual similarity within the stimulus set. An * indicates a reliable difference between the two illumination conditions. There was also a reliable interaction for sensitivity between the factors of Illumination and Experiment between the Shadow condition of Experiments 1 and 2, the same interaction between the Shadow and the WhiteShadow conditions in experiment 2, and a reliable response time difference between the Shadow and the NoShadow conditions in Experiment 2. Error bars show the normalized within-subject S.E.

stimuli, recognition performance was relatively fast and accurate. Second, when the direction of illumination was changed between the study and test images, there was a cost in the speed of recognition (although no cost in sensitivity). Note that the difference between the Change and NoChange conditions is unlikely to be due to local image properties independent of shape, in that a response time cost was found only for same trials, when a shape match is possible. Indeed, although not a reliable effect, the pattern obtained for different trials is actually in the opposite direction to that obtained for same trials. Thus, there is some evidence to support our attribution of this response time cost to the representation of the effects of illumination on the visual memory, an interpretation consistent with the image-based framework. On the other hand, it has been argued that response time effects without concurrent accuracy effects are not diagnostic with regard to the shape representations underlying performance in recognition experiments [41,42]. As discussed earlier, the use of objects that were qualitatively unique provides a more conservative test of the edge-based versus image-based cases. Consequently, the fact that we obtained a difference only in response times is not that surprising and there remains the possibility that this difference is due to local changes in the image independent of their status as shadows.

## 3. Experiment 2

The objects used in Experiment 1 were explicitly chosen so as to maximize the observers' reliance on shape information for object recognition. In particular, each of the six objects was qualitatively unique in terms of the shape of its central volume and the shapes of the attached parts. Moreover, the configuration of these parts was also unique (for one possible model of qualitative shape differences, see [25]). Given these extreme differences in shape, it is not that surprising that we obtained a reliable cost only in the response times across a change in illumination. Consider that Tarr and Bülthoff [21] argued that sensitivity to changes in viewpoint may be thought of as a continuum bounded by extreme invariance to extreme dependence. In their view, the most significant factor in determining viewpoint sensitivity is the difficulty of the discrimination, ranging from categorical judgments based on qualitative differences to exemplar-specific identification judgments based on subtle quantitative differences. A similar argument may be made for illumination (as the similarity between targets increases, we expect increasing illumination dependence). Overall, then, we hypothesized that increasing the similarity between targets, thereby pushing the participants to make more subtle shape discriminations, would produce both the response time difference seen in Experiment 1 and a concurrent sensitivity cost—the latter being somewhat more diagnostic with regard to representational issues [41,42].

### *3.1. Shadow condition*

Experiment 2 was designed to test this hypothesis by increasing the observers' reliance on configural information in the same sequential-matching task used in Experiment 1. Along with the original six objects, we included six new objects that were pairwise similar in terms of the shapes of the parts to those used in Experiment 1 (Fig. 2, bottom panel). This manipulation resulted in a discrimination task in which qualitative shape information alone was insufficient for discriminating between the members of the recognition set. Rather, it was assumed that the observers would rely more on configural information encoded in an image-based format. Thus, for a given image property, for example, the effects of illumination such as shading or shadows, we now predict larger costs in both the speed and the accuracy of recognition with a change in illumination direction.

### *3.2. NoShadow condition*

A second condition of Experiment 2 was designed to investigate which effects of illumination are responsible for the performance costs obtained when there is a change in illumination direction. Specifically, as noted in the image-based case, both informative shading and cast shadow markings may be represented in the visual memory. To dissociate these two effects, we re-rendered the combined set of twelve objects under the same lighting conditions, but without cast shadows (Fig. 5). Because all the other effects of illumination were identical with the Shadow condition, the NoShadow condition tested whether performance costs are due primarily to the relatively large changes in the image that are produced by cast shadows. Thus, if we obtain similar performance costs in the absence of cast shadows, we can infer that it is primarily surface shading that is encoded in visual memory. Alternatively, if we obtain smaller or no performance costs, we can infer that cast

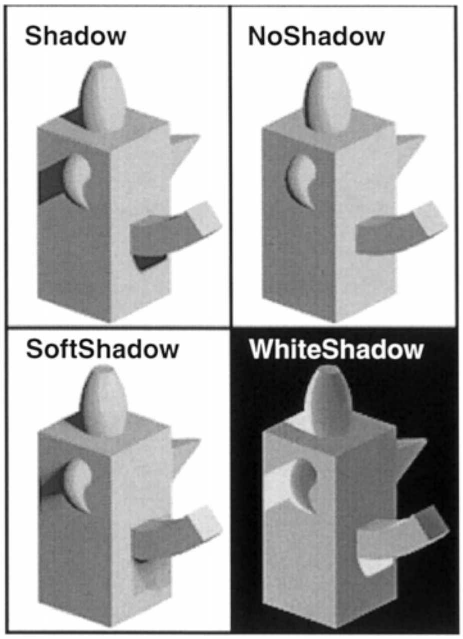

Fig. 5. An example of the Shadow, NoShadow, SoftShadow, and WhiteShadow conditions for one object. The top left image was rendered with sharp cast shadows; the bottom left image was rendered with soft cast shadows (having penumbrae); the top right image was rendered without shadows; and the bottom right image was rendered with sharp cast shadows and the contrast was inverted.

shadow markings are encoded in visual memory (as well as possibly surface shading).

Interestingly, a second consequence of removing the cast shadows is that the photometric similarity between the objects is increased. That is, while we have not altered the similarity between the shapes of the objects, we have made the images of the objects more similar. Thus, if the effects of illumination are encoded in the visual memory, and, in particular, are used to help distinguish between three-dimensional shapes in absence of other geometrical features (e.g. as in discriminating between cohorts in Fig. 2), then we would expect an overall decrease in performance regardless of any change in illumination. Such a finding would provide one clue as to why the visual system might encode the effects of illumination despite the variability inherent in this information. Specifically, the effects of illumination provide information regarding the three-dimensional shapes of objects (and, in particular, unfamiliar objects such as those used here).

### *3.3. WhiteShadow condition*

A control experiment for the Shadow condition of Experiment 2 was also run. This control was designed to investigate the origin of any performance costs in the Shadow condition. Specifically, a decrement in performance with a change in the direction of illumination might arise for two very different reasons. One possibility is that the performance costs might result from changes in illumination and shadows *qua* shadows—consistent with the idea that observers encode information about illumination in visual memory. A second possibility is that such costs are simply due to a comparison between identical versus changed images—a common and somewhat atheoretical result. To dissociate these two effects, we took the twelve objects from the Shadow condition and inverted the brightness for each object (Fig. 5).

The motivation behind this control is that inverting the brightness produces an image in which all of the contrast boundaries and objective image similarities are preserved from the original Shadow condition, but the shadows, now white, look less like shadows [12]. We hypothesized that brightness inverted images would be perceptually less similar to one another, because white shadows would be interpreted more as surface markings and less as shadows. As such we predicted a larger cost with white shadows when the lighting direction was changed relative to a constant lighting direction. We believe that this condition offers the best control possible for equating stimuli across experiments, but rendering the shadows less ‘shadow-like’. For example, relocating the shadows so that they are impossible and therefore less ‘shadow-like’ would produce new contrast edges and changes in the objective similarity between the images.

### *3.4. Method*

#### *3.4.1. Participants*

Seventy-six Yale undergraduates participated in return for course credit or pay: 32 participated in the Shadow condition and 44 participated in the NoShadow condition. An additional 34 Brown undergraduates participated in the WhiteShadow condition for pay. None of the participants had seen the stimuli prior to the experiment.

#### *3.4.2. Materials*

Along with the objects used in Experiment 1, six new objects that were pairwise similar to the original six objects were introduced (Fig. 2, bottom panel). Each of these new objects contained the same central volume and attached parts as one of the original objects, but was distinct from the original in terms of the configuration of the attached parts, i.e. location and orientation. Thus, in terms of local shape, each object was no longer qualitatively unique—objects could only be discriminated from their ‘cohorts’ by attending to the spatial relationships between the parts. All twelve of the objects shown in Fig. 2 were used in the Shadow, NoShadow, and WhiteShadow conditions. As shown in Figs. 2 and 5, objects in the Shadow condition had sharp cast shadows as in Experiment 1. Ray-traced shadows were not computed for the NoShadow condition, and thus produced images identical to the Shadow condition, but without cast shadows (Fig. 5). Finally, the images from the Shadow condition were contrast inverted (using a DeBabelizer, Equilibrium, Sausalito, CA) for the WhiteShadow condition. The mask used in the Shadow condition was rendered with sharp cast shadows, the mask used in the NoShadow condition was rendered without cast shadows, and the mask used in the WhiteShadow condition was a contrast inverted version of the mask from the Shadow condition.

#### *3.4.3. Design and procedure*

The same sequential-matching paradigm used in Experiment 1 was used for all of the conditions of Experiment 2. Only the number and composition of the trials was changed. Participants ran 288 test trials in which the 12 objects from Fig. 2 were shown in 2 illumination directions as described in Experiment 1. The objects appeared equally often in each illumination condition (12 times each) and, as in Experiment 1, the trials were divided evenly between same and different object pairs and the NoChange and Change in illumination conditions. Three shadow conditions were run between participants: a Shadow condition similar to that used in Experiment 1, in which the objects were rendered with sharp cast shadows; a NoShadow condition, in which the objects were rendered without cast shadows; and a WhiteShadow condition in which the contrast was inverted for the Shadow condition images, thereby creating somewhat less realistic looking 'white shadows'[4]. Other than the manipulation of the shadows, all other aspects of these three conditions were identical.

### 3.5. Results and discussion

Data analyses for all conditions were as described in Experiment 1.

#### 3.5.1. Shadow condition

We computed a sensitivity measure, *d'*, from the correct responses for same trials (hits) and the incorrect responses for different trials (false alarms) for each illumination condition: NoChange, 2.64; Change, 2.42. This is shown in Fig. 4, where a change in the direction of illumination produced a reliable difference in the ability of observers to judge that two images are the same object, $F(1, 31) = 11.0$, $P < 0.005$.

We also computed the mean response times for same and different trials for the NoChange and Change in illumination conditions (same/different): NoChange, 848 ms/827 ms; Change, 873 ms/838 ms. As shown in Fig. 4, a change in the direction of illumination produced a reliable 25 ms cost in judging that two images are the same object, $F(1, 31) = 8.90$, $P < 0.01$.

To compare these results with those of experiment 1, we ran ANOVA's on sensitivity and on same trial response times with two factors: Illumination Change, a within-subjects factor; and Experiment, a between-subjects factor. For sensitivity the main effect of Illumination Change was not reliable, $F < 1$, the main effect of Experiment was reliable, $F(1, 62) = 33.0$, $P < 0.001$, and the interaction was reliable, $F(1, 62) = 5.14$, $P < 0.05$[5]. Thus, it appears that our prediction of a sensitivity cost with the increasingly subtle shape discriminations used in Experiment 2 was confirmed. Moreover, we obtained the same difference in response times in both Experiments 1 and 2: the main effect of Illumination Change was reliable, $F(1, 62) = 17.5$, $P < 0.001$, the main effect of Experiment was not reliable, $F(1, 62) = 1.46$, ns (although response times were slower in Experiment 2), and the interaction was not reliable, $F < 1$.

The results of the Shadow condition of Experiment 2 replicates and extends the results of Experiment 1. First, as expected based on the higher shape similarity between the stimuli, recognition performance was relatively slower and less accurate. Second, we replicated the response time cost for a change in illumination (almost identical in magnitude to that obtained in Experiment 1). Third, in contrast to Experiment 1, we also found a sensitivity cost for a change in illumination—presumably because of the increased difficulty of the object discrimination used here. These sensitivity results provide stronger support for the idea that short-term object representations encode information about shading and cast shadows and, furthermore, suggest that there is an increasing reliance on image-based representations as shape discriminations become more difficult.

#### 3.5.2. NoShadow condition

We computed a sensitivity measure, *d'*, from the correct responses for same trials (hits) and the incorrect responses for different trials (false alarms) for each illumination condition: NoChange, 2.43; Change, 2.35. This is shown in Fig. 4, where a change in the direction of illumination did not produce a reliable difference in the ability of the observers to judge that two images are the same object, $F(1, 43) = 1.59$, ns.

We also computed the mean response times for same and different trials for the NoChange and Change in illumination conditions (same/different): NoChange, 952 ms/945 ms; Change, 967 ms/950 ms. As shown in Fig. 4, a change in the direction of illumination did not produce a reliable cost in judging that two images are the same object, $F(1, 43) = 2.61$, $P = 0.11$.

To compare these results with those from the Shadow condition, we ran ANOVA's on sensitivity and on same trial response times with two factors: Illumination Change, a within-subjects factor, and Experiment, a between-subjects factor. For *d'* the main effect of Illumination Change was reliable, $F(1, 74) = 8.90$, $P < 0.005$, the main effect of Experiment was not reliable, $F(1, 74) = 1.26$, ns (although the sensitivity was lower in the NoShadow condition), and the interaction was not reliable, $F(1, 74) = 2.07$, ns. For the response times the main effect of Illumination Change was reliable, $F(1, 74) = 9.76$, $P < 0.005$, the main effect of Experiment was reliable, $F(1, 74) = 4.05$, $P < 0.05$, and the interaction was not reliable, $F(1, 74) = 1.23$, ns.

The results of the NoShadow condition of Experiment 2 provide some constraints on our earlier findings. We found no difference in either the sensitivity or the response times across an illumination change when cast shadows were absent from the images. Thus, it appears

[4] To examine whether more realistic lighting conditions affect recognition performance we also ran a SoftShadow condition with 32 new participants. This condition was identical to the Shadow condition but with objects rendered with soft cast shadows (see [12] for a discussion of this issue. These images were created by rendering the shadows using an area light and the shadow boundaries were anti-aliased, producing smoothly changing realistic penumbrae.

[5] In the SoftShadow condition we obtained a reliable sensitivity difference comparable to that found for the Shadow condition: NoChange, 2.19; Change, 2.00; $F(1, 31) = 5.22$, $P < 0.05$; but a response time difference that was somewhat smaller and not reliable: NoChange: 787 ms; Change, 793 ms; $F(1, 31) = 1.95$, ns. There appears to a speed–accuracy trade off in this condition relative to the Shadow condition in that the overall sensitivity was lower but the response times were faster. At present we have no way of accounting for these differences.

as if the primary cause of illumination sensitivity is cast shadows. One interpretation of these results is that the shading variation is discounted early in recognition or that the visual system is able to extract useful information from the shading without any corresponding cost. In contrast, cast shadows are apparently not discounted, and as such, lead to some cost in performance when changed from one viewing episode to another. However, the NoShadow condition provides a clear reason why this should be so—without the presence of cast shadows, recognition was dramatically slower and somewhat less accurate. Thus, there is a compelling reason why the visual system encodes information about cast shadows—it provides useful information about three-dimensional structure that facilitates recognition (at least for novel objects such as those used here; in contrast, Braje et al. [36] found that cast shadows did not help recognition for faces, a known class).

### *3.5.3. WhiteShadow condition*

We computed a sensitivity measure, *d'*, from the correct responses for same trials (hits) and the incorrect responses for different trials (false alarms) for each illumination condition: NoChange, 2.37; Change, 1.91. This is shown in Fig. 4, where a change in the direction of illumination produced a reliable difference in the ability of the observers to judge that two images are the same object, $F(1, 33) = 25.3$, $P < 0.001$.

We also computed mean response times for same and different trials for the NoChange and Change in illumination conditions (same/different): NoChange, 844 ms/842 ms; Change, 869 ms/853 ms. As shown in Fig. 4, a change in the direction of illumination produced a reliable 25 ms cost in judging that two images are the same object, $F(1, 33) = 21.9$, $P < 0.001$.

To compare these results with those of the Shadow condition, we ran ANOVA's on sensitivity and on same trial response times with two factors: Illumination Change, a within-subjects factor, and Condition, a between-subjects factor. For sensitivity the main effect of Illumination Change was reliable, $F(1, 64) = 36.3$, $P < 0.001$, the main effect of Condition was reliable, $F(1, 64) = 11.4$, $P < 0.001$, and, critically, the interaction was reliable, $F(1, 64) = 4.33$, $P < 0.05$. Thus, our prediction of a larger sensitivity cost with white shadows (which we hoped would make shadows less 'shadow-like' and, therefore, more perceptually salient) was confirmed. Interestingly, we obtained almost an identical difference in the response times in the Shadow and WhiteShadow conditions: the main effect of Illumination Change was reliable, $F(1, 64) = 26.1$, $P < 0.001$, the main effect of Condition and the interaction were not reliable, both $F < 1$.

The results of the WhiteShadow condition of Experiment 2 offer important confirming evidence for our interpretation of Experiment 1 and the Shadow condition of Experiment 2. First, as predicted based on the less 'shadow-like' nature of white shadows [12], sensitivity costs were actually larger for white shadows as compared to black shadows across a change in illumination. Second, we replicated the response time cost for a change in illumination (almost identical in magnitude to that obtained in Experiment 1 and the Shadow condition). Thus, we have support for the claim that the sensitivity costs observed for a change in illumination in the Shadow condition are due to shadows *qua* shadows and not simply their presence in the image as local brightness variation. Such results support our claim that short-term object representations encode information about shading and cast shadows. How to interpret the response time costs, however, is less clear. Such costs were constant across the simpler shape discrimination used in Experiment 1, and the Shadow and WhiteShadow conditions of Experiment 2. Therefore, the response time differences may be associated more with mechanisms that normalize over local changes in the image, i.e. processing considerations, and less with specific properties of shape representations—an interpretation that is also consistent with the longer overall response times seen in the NoShadow condition (where the images became more similar).

## 4. Experiment 3

Experiments 1 and 2 used a sequential-matching paradigm that has some limitations—in particular, the technique assesses only short-term visual memory, while object recognition is typically based on long-term visual memory. In Experiment 3 we choose to use a naming paradigm in which observers first learned and practised the names of novel objects in one illumination and then had to name the same objects in a new illumination. Thus, we could investigate illumination sensitivity using a technique similar to the 'practice/surprise' technique introduced by Tarr and Pinker [43] in their study of orientation sensitivity. A second manipulation was introduced in Experiment 3—we manipulated the viewpoints of the objects between study and test. We were specifically interested in two alternative hypotheses regarding the interaction between familiarity of illumination and familiarity of viewpoint. First, if information about illumination is encoded in terms of its effects in the image, then any cost associated with a change in illumination should not generalize to new, unfamiliar viewpoints (which produce images that are likely to differ from those produced by the known viewpoints). In other words, we would not expect illumination sensitivity for viewpoints that have never been shown. Second, if information about illumination is encoded in terms of 'higher-level' descriptions, for example, a scene model or a set of basis functions (in Section 5 we will discuss two approaches to this problem, the illumination subspace model, first proposed by Shashua [29] (see also, [30]) and expanded by Hallinan [28,27], and the illumination cone model, a somewhat more realistic model that includes attached shadows proposed by Belhumeur and Kriegman [26]), then any cost associated with a change in illumination should, surprisingly, generalize to new, unfamiliar viewpoints—that is, viewpoints for which no illumination direction is familiar.

## *4.1. Method*

### *4.1.1. Participants*

Twenty members of the Max-Planck Institute in Tübingen, Germany participated in return for pay. The majority of participants were either students or postdocs and their ages ranged from approximately 18–26 years of age. None of the participants had seen the stimuli prior to the experiment.

### *4.1.2. Materials*

The objects used in Experiment 3 were identical to those used in the Shadow condition of Experiment 2. However, in addition to the 'canonical' viewpoint shown in Fig. 2, the objects were rendered in new viewpoints spaced at 30° intervals around the central vertical axis of the central volume of each object. These viewpoints were generated by rotating the object in depth while holding the lighting direction and the position of the viewer constant. Objects in each of these 12 viewpoints were rendered separately in the same two lighting directions used in Experiments 1 and 2, thereby producing 24 images per object.

### *4.1.3. Design and procedure*

A naming paradigm was used in Experiment 3. In the training phase participants were trained to associate nonsense names ('kip, kal, kef, kor, kym, and kug') with 6 'target' objects. Three target objects were selected from the 6 objects shown in the top panel of Fig. 2 and three target objects were selected from the 6 objects shown in the bottom panel of Fig. 2. Objects were selected so that each of the targets contained an unique central volume. Three targets were illuminated from one lighting direction and the other three were illuminated from the opposite lighting direction (for details see Section 2.1). Each object was seen under only one illumination direction, the 'Familiar' illumination. Participants were taught to recognize the targets in a series of 90 trials (15 per target object). On each trial a target object was shown in its canonical viewpoint along with its name written below the object. The participants were free to study each object as long as they wished and then pressed a key (one of 6) labeled with the appropriate name. The order of the trials was randomized for each participant.

Following training, in the practice phase the participants practised recognizing the target objects shown in the familiar illumination at the canonical viewpoint and at the 210° viewpoint. In addition to the 6 targets, participants were shown the 6 remaining unnamed objects as distractors and were told to respond 'none-of-the-above' for these objects. Thus, there were 7 responses, the 6 target keys and the 1 distractor key. Three distractors were illuminated from one lighting direction and the other three distractors were illuminated from the opposite lighting direction with the constraint that each distractor was illuminated from the opposite direction to its target cohort. The practice phase was organized into blocks of 96 trials composed of each target appearing 6 times in each of the 2 viewpoints and each distractor appearing 2 times in each of the 2 viewpoints—this resulted in 75% targets and 25% distractors. The order of the trials was randomized within each block and feedback in the form of a beep was provided for incorrect responses and trials in which the participants did not respond within 7500 ms. The participants were given extensive practice recognizing these targets over a period of two days. On the first day, 6 practice blocks were run and on the second day 2 additional practice blocks were run (for a total of 768 trials).

On the second day, following practice, in the surprise phase participants recognized the now-familiar objects from new viewpoints and new illumination directions. Both target and distractor objects were shown at 12 viewpoints in 30° increments around the vertical axis and in both their Familiar and Unfamiliar illuminations. The surprise phase was organized into one block of 576 trials composed of each target appearing 6 times in each of the 12 viewpoints and each distractor appearing 2 times in each of the 12 viewpoints—again a 75–25% split. The order of the trials was randomized within the block and feedback was provided (as discussed below, this feedback may have prompted participants to learn the familiar objects in the new illumination direction during the test session).

## *4.2. Results and discussion*

Only the results from the surprise phase will be considered—the training and practice phases being used solely for familiarizing the participants with a specific illumination and specific views for each object. For the purposes of computing the mean response times, the distractor trials, the incorrect responses, and the responses over 7500 ms were discarded. No adjustments were made to correct for outliers in that the response times were normally distributed. The mean response times were computed for target trials for each

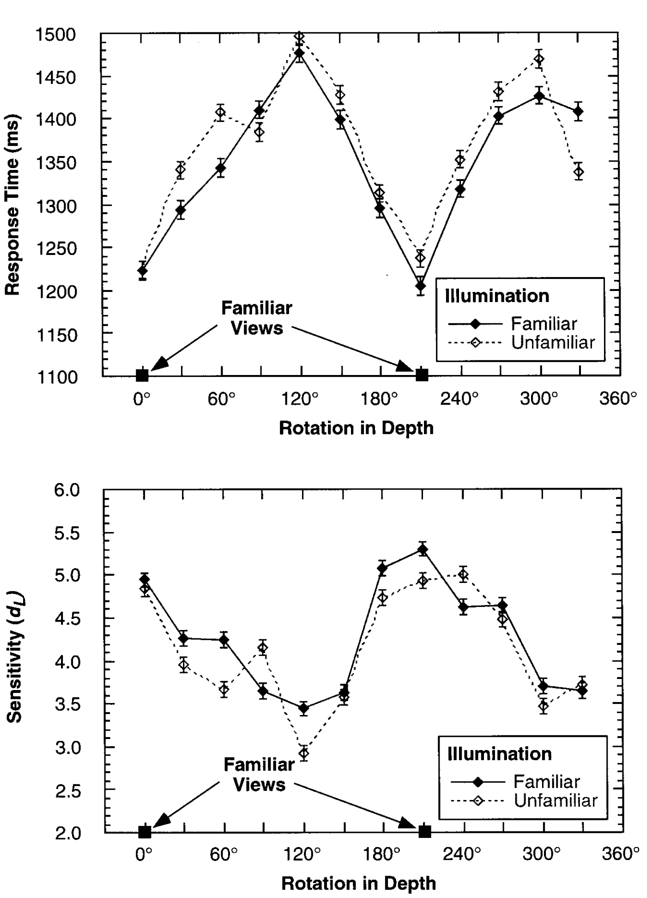

Fig. 6. Experiment 3. Mean response times for correct responses for target trials and mean sensitivity ($d_L$) across the Familiar and Unfamiliar illumination and the Familiar and Unfamiliar views conditions. Sensitivity can be measured in this experiment because the naming task includes distractor objects for which a name response constitutes a false alarm. For a discussion of this measure of sensitivity [44]. Error bars show the normalized within-subject S.E. for the effect of illumination familiarity.

viewpoint and for the Familiar and Unfamiliar illumination conditions (Fig. 6). The mean response times collapsed over viewpoint were computed for the Familiar illumination condition, 1349 ms, and the Unfamiliar illumination condition, 1368 ms. Recognizing a familiar object in an unfamiliar illumination resulted in a 19 ms cost in naming a target and recognizing a familiar object in an unfamiliar viewpoint resulted in an approximately linear increase in response time with increasing distance from a familiar viewpoint (for a discussion of similar 'multiple-views' performance patterns, see [20]). To investigate how these effects interacted, an ANOVA was run using Familiar/Unfamiliar Illumination and Viewpoint as within-subject factors. There was no reliable main effect for Illumination, $F(1, 19) = 2.51$, $P = 0.13$, a reliable main effect for Viewpoint, $F(11, 209) = 12.7$, $P < 0.001$, and no reliable interaction, $F < 1$. Although the effect of Illumination was not reliable when computed over participants, the same ANOVA computed over items revealed a reliable main effect for Illumination, $F(1, 5) = 6.29$, $P < 0.05$ (most likely because the variance between participants was much larger than the variance between items), a reliable main effect for Viewpoint, $F(11, 55) = 4.16$, $P < 0.001$, and no reliable interaction, $F < 1$. We also computed a sensitivity measure, $d_L$, based on the fact that the naming task includes target objects for which a name response may be considered a hit and distractor objects for which a name response may be considered a false alarm. Note that $d_L$ is functionally equivalent to $d'$, but is computed using logistic distributions [44]. We used $d_L$ rather than $d'$ because this measure is more stable for high hit rates and small false alarm rates. Mean sensitivity collapsed over viewpoint was computed for the Familiar illumination condition, 4.26, and the Unfamiliar illumination condition, 4.12. As shown in Fig. 6, recognizing a familiar object in an unfamiliar illumination resulted in a small sensitivity cost in discriminating targets from distractors and recognizing a familiar object in an unfamiliar viewpoint resulted in an approximately linear decrease in sensitivity with increasing distance from a familiar viewpoint (again see [20]). To investigate how these effects interacted, we again ran an Illumination × Viewpoint ANOVA. There was no reliable main effect for Illumination, $F(1, 19) = 1.32$, ns (some learning may have occurred during the test session. An analysis on the first half of the trials for each participant revealed much larger costs for the Unfamiliar illumination condition as compared to the Familiar illumination condition. For example, at 0° the response time difference was 60 ms and the $d_L$ difference was 0.42—the fact that these costs diminished rapidly during testing may reflect how readily observers are able to learn new illumination conditions for highly familiar viewpoints), a reliable main effect for Viewpoint, $F(11, 209) = 12.8$, $P < 0.001$, and no reliable interaction, $F(11, 209) = 1.35$, ns.

For both response times and sensitivity, we also investigated the results in terms of a qualitative interaction between Illumination and Viewpoint by recoding each view as either simply familiar or unfamiliar. This analysis again showed no hint of a reliable interaction for either response times or sensitivity, $F < 1$ in both cases.

The results of Experiment 3 further extend the results of Experiments 1 and 2. First, consistent with the image-based case in which the effects of illumination are thought to be encoded in the visual memory, we again obtained evidence for a decrement in recognition performance with a change in illumination direction. Importantly, unlike our earlier experiments, this illumination dependency cannot be explained by low-level sensory or even more transient short-term visual memories. Rather, illumination dependency in Experiment 3 can only be accounted for by the representation of the effects of illumination in long-term visual memory—in particular, because sensitivity to a change in illumination was obtained across relatively long intervals between study and test and because naming tasks are thought to rely on more global 'high-level' representations as opposed to local sensory codes. Second, consistent with the idea that we represent information about illumination sources, not simply its effects in the image, we found no interaction between familiarity of illumination direction and familiarity of viewpoint. Consider that the same cost for a change in illumination was obtained for both familiar and unfamiliar viewpoints. If the effects of illumination are encoded simply as they appear in the image, then any cost for a change in illumination should occur only at familiar viewpoints where the original effects of illumination were actually seen. In contrast, if the effects of illumination are encoded as a model of the scene or as a set of basis functions modeled by either a subspace [29,30,28,27] or a cone [26], then a cost for a change in illumination might be expected to occur for both familiar and unfamiliar viewpoints. Such was the case in Experiment 3. Thus, we can surmise that the effects of illumination are not simply preserved as they originally appear, but rather are modeled in some sense by high-level visual processes—a reasonable interpretation in light of the degree to which the presence of cast shadows facilitated recognition in Experiment 2. Indeed, the facilitation associated with cast shadows was far larger than the costs associated with variations in illumination found in Experiments 1 and 2. Thus, the benefits for object recognition inherent in representing the effects of illumination seem to outweigh the costs of including variability arising from shading and cast shadows.

## 5. General discussion

### *5.1. Illumination sensitivity*

We began by pointing out that a key problem in object recognition is that images of objects vary depending on the viewpoint and illumination conditions. When considered in the context of edge- and contour-based theories of recognition [25,2], the typical story has been that early on the visual system recovers invariant features, but discounts 'spurious' features, in particular, the effects of illumination in the image. When considered in the context of more recent image-based theories of recognition [3,5,6], however, it is possible that high-level object representations retain much of the information present in the input image [4]. Thus, from at least one paradigmatic perspective there are reasons to believe that variations in properties such as shading, shadows, and texture may influence object recognition behavior. Moreover, when we specifically focus on cast shadows from a computational perspective, there are reasons to believe that these particular effects of illumination are difficult to discount. As such, cast shadows may be included in longer-term object representations by default. This situation, however, is not all bad news. From the same computational perspective, the interactions between a light source, the casting object, and the receiving surface, all encrypted within cast shadows, provide valuable information regarding three-dimensional structure [12]. Therefore, the inclusion of the effects of illumination in longer-term object representations may actually be desirable, in particular, for novel objects such as those used here.

In our investigation of whether human object recognition is sensitive to changes in illumination, we found several key results:

- In the presence of cast shadows the participants were slower at matching objects when the lighting direction changed.
- In the presence of cast shadows the participants were slower and less sensitive at matching objects when the lighting direction changed and when the objects to be discriminated were similar.
- In the presence of white shadows the participants were slower and even less sensitive at matching objects when the lighting direction changed.
- In the absence of cast shadows the participants did not show the same costs in matching objects across changes in the lighting direction.
- In the absence of cast shadows the overall performance was slower and less sensitive relative to the same task with cast shadows.
- Objects familiar in a given lighting direction were named more slowly and with lower sensitivity when shown in an unfamiliar lighting direction.
- The cost for naming an object in an unfamiliar lighting direction was nearly equivalent for familiar and unfamiliar viewpoints.

This pattern of results over experiments is informative regarding both the processing and representation of the effects of illumination. Consider the constant increase in response times across changes in illumination whenever shadows were present. These costs seem best associated with the processing or normalization of shadows as local image features that were changing from image to image. On the other hand, the decrease in sensitivity across changes in illumination varied depending on the difficulty of the shape discrimination and the ability of the observer to account for the shadows as the effects of illumination (rather than as local image features). Thus, these costs seem best associated with the representation of shadows as shadows. As such, they provide some evidence for the representation of the effects of illumination in visual memory. More specifically, we offer four conclusions. First, both short-term visual memory (Experiments 1–2) and longer-term (Experiment 3) visual memory are affected by changes in illumination. Second, much of the cost arising from variation in lighting direction appears to be associated with the presence of cast shadows that may produce spurious edges or surfaces (Experiment 2). Third, while it may be true that the effects of illumination are difficult to discount in early vision, it also seems that they serve a useful function—that of disambiguating three-dimensional shapes. Fourth, with regard to the nature of the representation, we hypothesize that the effects of illumination are implicitly modeled, rather than simply retained as artifacts within the image (experiment 3). We discuss the first and second points in the next section. The usefulness of cast shadows and the nature of illumination representation are discussed in the two following sections.

### *5.2. Discounting the effects of illumination is hard*

Illumination variation poses a stiff computational challenge to object recognition. Recent studies have demonstrated that none of the traditional computer vision methods designed to extract putative illumination-invariant features, such as edge maps, intensity derivatives, or Gabor filter representations are sufficient to achieve acceptable illumination-invariant recognition as compared with human performance [45,8,46]. The difficulty lies in the fact that the sources of intensity variation—shadow, material, and shape—are confounded in the pattern of image intensities. Cast shadows and specularities are particularly problematic because the causes of their intensity changes are not local to the surface, and result in edges with an indirect and highly ambiguous relationship to the underlying shape.

Brain lesion studies also support the conclusion that the compensation for illumination is not a simple matter of appropriate early filtering processes. Warrington [47] found patients with right posterior lesions who had difficulties compensating for both view and illumination changes. In monkey studies, lesions of the anterior inferotemporal lobe severely reduce a monkey's capacity to recognize objects, including images of objects previously seen [48]. Lesions to parts of inferotemporal cortex and prestriate areas hamper a monkey's ability to generalize over size and illumination change. These results are consistent with the idea that the inferotemporal cortex is involved in the storage of object prototypes useful for generalizing across both view and lighting. In single unit recordings from the superior temporal sulcus in the temporal cortex of macaque monkey, Hietanen et al. [49] found some view-specific cells selective for faces that were sensitive to illumination changes. As a population (about 20 cells), however, the responses were fairly illumination invariant. These results again suggest that the inferotemporal cortex is involved in the representation of objects with some measure of invariance to view and lighting changes. Further, the single cell results are also consistent with this invariance being achieved through the distributed storage of both view-specific and illumination-specific information.

#### *5.2.1. Familiar versus unfamiliar object classes*

We have already pointed out the theoretical difficulties presented by cast shadows. One solution to the dilemma of local ambiguity is to identify cast shadows through top-down information available for familiar object classes, perhaps specified by an object prototype [47,7]. Moore and Cavanagh [16] studied object recognition for so-called 'Mooney' or two-tone images, in which all pixels with gray-levels below a fixed threshold are set to black, and those above are set to white. Mooney images exacerbate the confusion of the material, shape, and shadow causes to intensity change by removing all local cues to the identity of these edge types. They found that Mooney images from familiar objects classes, e.g. faces, are more easily recognized than Mooney images from unfamiliar classes, e.g. collections of three-dimensional volumes. Their interpretation is that there may be sufficient information to access the object class, and that this information can be used to resolve local ambiguities in the causes of the intensity changes.

Adini et al. [45] (see also [8,46]) studied recognition generalization to new views and illumination directions for inverted and upright faces. They found good generalization to novel images for upright faces, but generalization to novel views and illuminations was significantly worse for inverted faces. They interpreted this difference to be a consequence of the way in which the visual system deals with illumination and view change for the inverted and upright face object classes. In a study of the well-known face inversion effect [50], Johnson et al. [40] found that the cost in recognizing an inverted face was substantially reduced for faces illuminated from below, as compared with the standard case in which the face is illuminated from above.

Braje et al. [36] directly studied the effects of cast shadows for face recognition in a sequential-matching task similar to that used in Experiments 1 and 2. Consistent with the results of this paper, they found a cost to a change in illumination for face recognition; however, in contrast to Experiment 2, the presence of cast shadows impaired the overall recognition performance. At first, this result appears to contradict the idea that familiar classes should benefit more from cast shadows, in that cast shadows are less ambiguous. However, two factors suggest that a well-known class such as faces may not benefit from the presence of cast shadows. First, strong cast shadows may obscure critical details in complex images such as faces. Second, strong familiarity with the three-dimensional shape of well-known classes such as faces may reduce any potential benefit provided by cast shadows. In comparison, the objects used in the present experiments had much simpler images with far fewer subtle details required for discrimination and they had completely unknown three-dimensional shapes. Thus, the role of cast shadows in object recognition may vary with both the geometry and the familiarity of the class.

### 5.3. Cast shadows are useful

Earlier we suggested that our observers benefited from encoding some of the effects of illumination. Our conjecture is that while one source of this benefit clearly comes from shape-from-shading processes [9,10], another, less often considered source, is the information about the relative depth between an object part and the surface receiving the cast shadow (Experiment 2, Shadow versus NoShadow condition). Recent results by Kerstan et al. [12] reinforce the point that cast shadows provide valuable information about three-dimensional structures, for example, the spatial layout of a scene. They showed that a moving cast shadow was sufficient to induce apparent motion in depth of the casting object, even when the object itself was perfectly stationary relative to the background. Effectively utilizing cast shadow information for depth requires the solution of two problems. The visual system must determine: (1) whether the shadow region is in fact a cast shadow, rather than a material change (e.g. dark paint), or a change in surface orientation; (2) the correspondence between the shadow and its casting object. Exactly how this is, or even can be, done is not entirely clear. The identification of shadows probably involves the combination of weak low-level cues, such as shadow fuzziness and contrast-invariant darkening of the background (as evidenced by the differences in the Shadow and WhiteShadow conditions of Experiment 2), together with high-level processing of the sort discussed in the previous section.

### 5.4. The representation of the effects of illumination

In the longer-term naming task of Experiment 3, the participants showed the effect of illumination on both response time and sensitivity. The participants also showed a viewpoint effect that, interestingly, did not interact with costs due to illumination variation. The fact that there were costs for a change in illumination for viewpoints that participants had never seen suggests that illumination is not represented simply in terms of its effects on the image, but rather, is implicitly modeled with respect to the shape of a given exemplar. On the other hand, whatever model of lighting is used, it does seem that it does not apply to the entire scene: there was no cost to illumination change on different trials in Experiments 1 and 2.

Other evidence in the literature supports this hypothesis, suggesting that the visual system does not represent illumination direction with respect to the overall scene. Perception does not insist on global consistency for illumination direction for either depth or shape perception. For example, when determining an object's three-dimensional trajectory from its cast shadow, the visual system ignores inconsistent shadows from other objects in the scene [51]. In a pair of animations, several cubes were placed on the floor of a box. An egg-shaped object was made to move in a diagonal trajectory above the floor. In one condition, the light source moved to produce a shadow trajectory consistent with that resulting from a stationary light source illuminating an egg sliding on the floor. In the second condition, the light source moved in such a way as to produce a shadow consistent with a stationary light source illuminating an egg flying through the air. In both conditions, however, the egg had the same three-dimensional trajectory as well as identical two-dimensional trajectories in the image. Perception, however, ignores the information from the shadows cast by the stationary cubes that should have informed the visual system that the trajectories in the two animations were identical; instead, the egg appeared to slide along the floor in the first case, and fly above the floor in the second. (For a demonstration of this effect, see: http://www.mpik-tueb.mpg.de/projects/genlight/snake_w_box.mpg.) Studies of shape-from-shading also offer evidence that illumination direction is tied to individual objects [39]. For example, a hemispherical bump illuminated from below often appears as a concave surface illuminated from above; however, when the bump appears attached to a human face illuminated from below, it appears convex [39]. The reason is that the convexity of a face is not ambiguous, being a familiar surface; thus it appears illuminated from below, and the bump is interpreted consistently as a bump. If the bump is perceived as independently of the face, it is then more likely to appear concave, presumably because in this case, the visual system does not apply a globally consistent illumination direction to both the face and the bump. Taken together, these observations suggest that the representation of illumination may be tied to individual objects, or small portions of scenes, rather than to the scene as a whole.

#### 5.4.1. Computational models of illumination

We have already discussed the lack of success of the traditional computer vision approach to illumination variation—edge detection and/or early spatial filtering to illumination variation is inadequate. The fact that we obtain an illumination-specific cost to recognition argues against the kind of 'smart' edge-based representation described in the introduction. What are the alternatives? Another traditional approach in computer vision is to use independent and explicit knowledge of global light source direction to determine shape-from-shading through local filtering [52,53]. The fact that human vision does not seek global consistency argues against a general-purpose illumination direction estimator applicable to the scene as a whole. It does not, however, rule out explicit object-specific estimates of illumination direction. Whether to represent illumination explicitly or not is still a matter of debate in computer vision [27]. An explicit estimate of illumination parameters, can in principle, be useful for other tasks in addition to recognition. Furthermore, it provides the means for better generalization to novel illumination conditions [26,28,54,29,30]. So-called 'appearance-based' (a term typically used in the computer vision literature to refer to low-dimensional representations of the class of images of an object, e.g. achieved through principal components analysis, that do not explicitly distinguish shape, material, and lighting) models of object recognition avoid the explicit representation of illumination, and instead rely on stored illumination- and/or view-specific object models for recognition [55]. However, it is difficult to see how to extend these models to handle novel combinations of illumination and material changes.

One recent approach to dealing with illumination variation in object recognition is to use a linear model of illumination to derive illumination basis images for a specific view of an object [27,28,54,29,30]. These 'eigenimages' define the space over which the model can generalize to new illumination conditions (see [26] for a potentially more robust variation of this approach). The linear model seems to work well, despite the violations caused by cast shadows and specularities. When fitting the model to the image data, cast shadows and specularities can be treated as residuals. This kind of model leaves open the option of a secondary process to evaluate the residual. It is not difficult to see a connection between the view-specific and illumination-specific cells of the inferotemporal cortex and the basis images in this kind of computational approach. Further, the notion of residuals is what is required to identify cast shadows by means of a secondary top-down process following the preliminary identification of the object itself. Thus, we are left with a challenge—our current results suggest that we should not ignore the effects of illumination in theories of object recognition, yet we have little specific understanding of how the human visual system processes and represents shading and cast shadows. Hopefully, future psychophysical studies will expand on these issues taking into account the predictions of various well-specified and potentially powerful computational models of illumination representation.

## Acknowledgements

This research was supported by a grant from the Air Force Office of Scientific Research, contract number F49620-91-J-0169, and a Max-Planck Institute visitor's grant to MJT, and grants from NATO and the National Science Foundation, contract number SBR-9631682, to DK. We thank Alan Yuille for his thoughtful comments and discussions on many of the ideas in this paper, Scott Klemmer, James Rosoff, Alexa Rupertsberg, Jay Servidea, Amy Taxin, and KaRin Turner for their assistance in running the experiments and Scott Yu for creating the stimuli.

## References

[1] Biederman I, Ju G. Surface versus edge-based determinants of visual recognition. Cogn Psychol 1988;20:38–64.

[2] Marr D, Nishihara HK. Representation and recognition of the spatial organization of three-dimensional shapes. Proc R Soc London B 1978;200:269–94.

[3] Bülthoff HH, Edelman SY, Tarr MJ. How are three-dimensional objects represented in the brain? Cereb Cortex 1995;5(3):247–60.

[4] Edelman S. Representation, similarity, and the chorus of prototypes. Minds Mach 1995;5(1):45–68.

[5] Gauthier I, Tarr MJ. Orientation priming of novel shapes in the context of viewpoint-dependent recognition. Perception 1997;26:51–73.

[6] Poggio T, Edelman S. A network that learns to recognize three-dimensional objects. Nature 1990;343:263–6.

[7] Cavanagh P. What's up in top-down processing? In: Gorea A, editor. Representations of Vision: Trends and Tacit Assumptions in Vision Research. Cambridge, UK: Cambridge University Press, 1991:295–304.

[8] Moses Y, Adini Y, Ullman S. Face recognition: The problem of compensating for changes in the illumination direction. In: Eur, Stockholm Conf on Computer Vision. 1994:286–296.

[9] Horn, BKP. Obtaining shape from shading information. In: Winston PH, editor. The Psychology of Computer Vision. New York: McGraw-Hill, 1975:115–155.

[10] Langer MS, Zucker SW. Shape from shading on a cloudy day. J Opt Soc Am 1994;A 11(2):467–78.

[11] Koenderink JJ, van Doom AJ, Kappers AML. Surface perception in pictures. Percept Psychophys 1992;52(5):487–96.

[12] Kersten D, Knill DC, Mamassian P, Bülthoff I. Illusory motion from shadows. Nature 1996;379:31.

[13] Bülthoff, H.H. Shape from X: psychophysics and computation. In: Landy MS, Movshon JA, editors. Computational Models of Visual Processing. Cambridge, MA: MIT Press, 1991.

[14] Cavanagh P, Leclerc YG. Shape from shadows. J Exp Psychol Hum Percept Perform 1989;15(1):3–27.

[15] Bülthoff I, Kersten D, Bülthoff HH. General lighting can overcome accidental viewing. Assoc Res Vis Ophthalmol 1994;35:S1741.

[16] Moore C, Cavanagh P. Recovery of 3d volume from 2-tone images of novel objects. Cognition 1998 (in press).

[17] Bülthoff HH, Edelman S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc Natl Acad Sci USA 1992;89:60–4.

[18] Humphrey GK, Khan SC. Recognizing novel views of three-dimensional objects. Can J Psychol 1992;46:170–90.

[19] Jolicoeur P. The time to name disoriented natural objects. Mem Cogn 1985;13:289–303.

[20] Tarr MJ. Rotating objects to recognize them: a case study of the role of viewpoint dependency in the recognition of three-dimensional objects. Psychon Bull Rev 1995;2(1):55–82.

[21] Tarr MJ, Bülthoff HH. Is human object recognition better described by aeon-structural-descriptions or by multiple-views? J Exp Psychol Hum Percept Perform 1995;21(6):1494–505.

[22] Edelman S. Class similarity and viewpoint invariance in the recognition of 3D objects. Biol Cybern 1995;72:207–20.

[23] Hummel JE, Biederman I. Dynamic binding in a neural network for shape recognition. Psychol Rev 1992;99(3):480–517.

[24] Hummel JE, Stankiewicz BJ. An architecture for rapid, hierarchical structural description. In Inui T, McClelland J, editors. Attention and Performance vol. 15. Cambridge, MA: MIT Press, 1996:93–121.

[25] Biederman I. Recognition-by-components: a theory of human image understanding. Psychol Rev 1987;94:115–47.

[26] Belhumeur P, Kriegman D. What is the set of images of an object under all possible lighting conditions? In: IEEE Conf. on Computer Vision and Pattern Recognition. San Francisco, CA, 1996:270–277.

[27] Epstein R, Hallinan PW, Yuille, AL. 5 ± Eigenimages suffice: an empirical investigation of low-dimensional lighting models. In: IEEE Workshop on Physics-Based Modeling in Computer Vision. Boston, MA, 1995:108–11.

[28] Hallinan, PW. A low-dimensional lighting representation of human faces for arbitrary lighting conditions. In: IEEE Conf. on Computer Vision and Pattern Recognition. Seattle, 1994:995–999.

[29] Shashua A. Illumination and view position in 3D visual recognition. In: Moody JE, Hanson SJ, Lippmann RP, editors. Advances in Neural Information Processing Systems vol.4. San Mateo, CA: Morgan Kaufmann, 1992:404–411.

[30] Shashua A. On photometric issues in 3D visual recognition from a single 2D image. Int J Comput Vis 1996;21:99–122.

[31] Biederman I, Gerhardstein PC. Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. J Exp Psychol Hum Percept Perform 1993;19(6):1162–82.

[32] Tarr MJ, Bülthoff HH, Zabinski M, Blanz V. To what extent do unique parts influence recognition across changes in viewpoint? Psychol Sci 1997;8(4):282–9.

[33] Ellis R, Allport DA. Multiple levels of representation for visual objects: a behavioural study. In: Cohn AG, Thomas JR, editors. Artificial Intelligence and its Applications. New York: Wiley, 1986:245–247.

[34] Biederman I, Cooper EE. Size invariance in visual object priming. J Exp Psychol Hum Percept Perform 1992;18(1):121–33.

[35] Biederman I, Cooper EE. Evidence for complete translational and reflectional invariance in visual object priming. Perception 1991;20:585–93.

[36] Braje WL, Kersten D, Tarr MJ, Troje NF. Illumination and shadows influence face recognition. Invest Ophthalmol Vis Sci 1996;37:S176.

[37] Goodale MA, Jakobson LS, Keillor MM. Differences in the visual control of pantomimed and natural grasping movements. Neuropsychologia 1994;32(10):1159–78.

[38] Brewster D. On the optical illusion of the conversion of cameos into intaglios and of intaglios into cameos, with an account of other analogous phenomena. Edinb J Sci 1826;4:99–108.

[39] Ramachandran VS. Perceiving shape from shading. Sci Am 1988;256(6):76–83.

[40] Johnston A, Hill H, Carman N. Recognising faces: effects of lighting direction, inversion, and brightness reversal. Perception 1992;21:365–75.

[41] Edelman S, Bülthoff HH. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vis Res 1992;32(12):2385–400.

[42] Bülthoff HH, Edelman S. Evaluating object recognition theories by computer graphics psychophysics. In: Poggio TA, Glaser DA, editors. Exploring Brain Functions: Models in Neuroscience. New York, NY: Wiley, 1993:139–164.

[43] Tarr MJ, Pinker S. Mental rotation and orientation-dependence in shape recognition. Cogn Psychol 1989;21(28):233–82.

[44] Snodgrass JG, Corwin J. Pragmatics of measuring recognition memory: Applications to dementia and amnesia. J Exp Psychol Gen 1988;117:34–50.

[45] Adini Y, Moses Y, Ullman S. Face recognition: the problem of compensating for changes in illumination direction. Technical Report. The Weizmann Institute of Science, 1995.

[46] Moses Y, Ullman S, Edelman S. Generalization to novel images in upright and inverted faces. Perception 1996;25:443–62.

[47] Warrington EK. Neuropsychological studies of object recognition. Phil Trans R Soc London 1982;B 298:15–33.

[48] Weiskrantz L. Visual prototypes, memory, and the inferotemporal lobe. In Mishkin EI, editor. Vision, Memory and the Temporal Lobe. New York: Elsvier, 1990:13–28.

[49] Hietanen JK, Perrett DI, Oram MW, Benson PJ, Dittrich WH. The effects of lighting conditions on responses of cells selective for face views in the macaque temporal cortex. Exp Brain Res 1992;89:151–71.

[50] Yin RK. Looking at upside-down faces. J Exp Psychol 1969;81(1):141–5.

[51] Kersten D, Mamassian P, Knill DC. Moving cast shadows induce apparent motion in depth. Perception 1997;26(2):171–92.

[52] Kersten D, O'Toole AJ, Sereno ME, Knill DC, Anderson JA. Associative learning of scene parameters from images. Appl Opt 1987;26:4999–5006.

[53] Pentland AP. Linear shape from shading. Int J Comput Vis 1990;4:153–62.

[54] Hallinan, PW. A deformable model for face recognition under arbitrary lighting conditions. Unpublished PhD thesis, Division of Applied Sciences, Harvard University, 1995.

[55] Murase H, Nayar S. Visual learning and recognition of 3-D objects from appearance. Int J Comput Vis 1995;14:5–24.