High-level Vision as Statistical Inference

Daniel Kersten

SHORT TITLE: Vision as Statistical Inference

kersten@tc.umn.edu

Department of Psychology, University of Minnesota, 75 East River Road,Minneapolis, MN, 55455.

Kersten, D. (1999) The New Cognitive Neurosciences--2nd Edition, Gazzaniga, M.S. (Ed.), MIT Press.

ABSTRACT

Human vision is remarkably versatile and reliable, despite the fact that retinal image information is noisy, ambiguous, and confounds the properties of objects that are useful. By treating vision as a problem of statistical inference, three classes of constraints can be identified: the visual task, prior knowledge of scene structure independent of the image, and the relationship between image structure and task requirements. By considering the visual system as an organ for statistical inference, we can test whether and how it uses these constraints. This strategy is illustrated for two high-level visual functions: depth-from-cast-shadows and viewpoint compensation in 3-D object recognition.

An object's relative depth can be determined from its cast shadow, even when local image information doesn't uniquely specify shadow edges, and global information doesn't determine where the light source is. What information enables a unique estimate of depth from shadows? This chapter shows how the visual task, prior assumptions on light movement and material properties, and local image cues constrain the perception of depth from shadows.

A 3D object can be recognized from views never seen before, despite the fact that depth information about shape is lost due to projection on to the retina. How does human recognition compensate for variations in viewpoint? By designing a simple recognition task for which optimal statistical decisions are computable, human performance can be normalized with respect to the information in the task, leaving remaining differences diagnostic of brain mechanisms.

High-level vision is often divided into two primary functions: object recognition and localization. Although these visual functions have quite different processing demands, they are linked by a common framework of statistical inference. This chapter takes seriously the idea that vision consists of brain processes for statistical decisions and estimation (Kersten, 1990; Yuille and Bülthoff, 1996). Perception as inference has a long history; however, it is with the advent of computer vision that we have begun to understand the inherent complexity of visual inference from natural images. The challenge has spurred the development and application of theoretical tools for modeling visual inference (Clark and Yuille, 1990; Knill and Richards, 1996). The problem of vision is both geometrical and photometrical: the depth dimension is lost due to projection onto the retina, and information about the geometry of objects gets entangled with photometric information about object material and illumination. Because image intensity at a point is a function of object shapes, materials, illumination, and viewpoint, information about the world is encrypted in the pattern of image intensities (Figure 1).

Much of our knowledge of the visual system has come through an analysis of early stages of processing in which we try to understand how local contours (defined by intensity, color, texture, disparity or motion) are grouped to define objects. Figure 2 shows how the local constraints of edge colinearity and transparency determine how contours are grouped, and as a result one sees either overlapping ape faces, or non-overlapping human faces. Research in computer vision has shown, however, that edge detection and object segmentation from natural images is a harder problem. Even apart from ubiquitous image and neural noise, the response of an optimally tuned oriented spatial filter (e.g. simple cell in visual cortex) does not uniquely determine whether the corresponding edge in the scene is due to a shadow, specularity, or a change in depth, orientation or material (Figure 3). Yet, such distinctions are crucial for visual function. Adaptive visual behavior depends on reliable decisions regarding object shape, material, and spatial relationships (Figure 1). Because of the inherent ambiguity in the eye's input regarding these scene properties, vision is sometimes said to be an ill-posed problem. In contrast, the brain has clearly

solved the problem--but how? The answers lie both in the nature of visual mechanisms, and in the theme of this chapter--the information that constrains visual decisions.

1. Natural constraints and visual decisions

One can identify three types of constraints that make reliable visual inference possible: the visual task, prior knowledge of scene structure independent of the image, and the relationship between image structure and task requirements.

Bayesian decision theory provides a precise language to model these constraints (Yuille and Bülthoff, 1996). We postpone discussion of the visual task, and suppose the image measurements, **I**, and the required scene parameters, **S**, useful for the task have been specified. The knowledge for visual inference is characterized by the posterior probability distribution, $P(\mathbf{S}|\mathbf{I})$ which models the probability of a scene description **S**, given the image data, **I**. By Bayes' rule, the posterior is:

$$P(\mathbf{S} \mid \mathbf{I}) = \frac{P(\mathbf{S})P(\mathbf{I} \mid \mathbf{S})}{P(\mathbf{I})} \qquad P(\mathbf{S})P(\mathbf{I} \mid \mathbf{S}) = P(\mathbf{S})P(\mathbf{I} - F(\mathbf{S}))$$

where $P(\mathbf{I})$ is fixed for a given image measurement.

$P(\mathbf{S})$ is the prior distribution modeling the scene. In theory, a prior scene model could be realized as an algorithm to produce samples of scenes, including objects, materials, illuminations, independent of the images that might result. In practice, we are limited to modeling subdomains such as surface smoothness, shape, contour, or material (Kersten, 1991; Poggio, Torre, and Koch, 1985; Sha'ashua and Ullman, 1988; Zucker and David, 1988), or specific object domains (Troje and Vetter, 1996). From the standpoint of inference, knowledge of prior constraints eliminates alternative image interpretations which are consistent with the image data. Later we'll how the assumption that light sources are usually above objects affects the perception of depth from cast shadows (Section 2).

$P(\mathbf{I} \mid \mathbf{S})$ is the likelihood of the image measurements given a scene description. The likelihood is determined by how images are formed--the image rendering problem of computer graphics, $\mathbf{I} = F(\mathbf{S})$. A common example of the likelihood constraint is that straight lines in the scene project to straight lines in the image. The likelihood also provides the tools for reducing ambiguity through cue integration (Landy et al., 1995). *A priori* knowledge of the scene would

seem to be required to develop an explicit model of the image. However, Bayes provides tools for learning representations of the image, bottom-up (e.g. Olshausen and Field, 1996; Zhu, Wu, and Mumford, 1997). Mumford (1995) has proposed that minimum description length encoding (formally equivalent to Bayes maximum a posteriori estimation) may provide a general means to discover world structure from images. Specific task requirements can also be used to discover useful image features (e.g. Belhumeur, Hespanha, and Kriegman, 1996). These image-based approaches are important because the problems are posed in a form closer to those of natural adaptation and development. But ultimately, the statistical structure of images derives from how images are formed from the scene.

Let's return to the issue of how knowledge of the visual task reduces ambiguity in visual inference.

Specifying the task--explicit and generic variables

Visual problems are often said to be ill-posed when there are more scene parameters to estimate than data. In this case, priors are essential to find unique solutions. However, for specific functional goals, such as visual tracking or face recognition, the number of parameters can be drastically reduced (Blake and Yuille, 1992). With a good representation, the prior is constant, and the decision can be made on the likelihood alone. One still faces the problem that image intensities confound *all* of the scene variables, both the irrelevant, and those required for the task. The relevant and irrelevant scene variables are called *explicit* and *generic*, respectively[1]. The general

---

[1]More generally, Bayesian decision theory softens the sharp distinction between explicit and generic variables by defining a loss function $L(S, \Sigma)$ which is the penalty for Σ, (the estimate of S) when the true scene parameter is S. Then the optimal decision minimizes the risk:

$$R(\Sigma_G, \Sigma_E) = \int L(S_G, \Sigma_G; S_E, \Sigma_E) P(S_E, S_G | I) dS_G dS_E$$

idea is that different visual tasks require a more explicit or precise representation of some scene parameters than others (Brainard and Freeman, 1994; Freeman, 1994; Yuille and Bülthoff, 1996; Figure 1, upper box). For example, object recognition relies on an estimation of shape, with viewpoint discounted (Section 3). But discounting is *not* ignoring, and one would like estimates of the scene which are insensitive to the generic variables. In fact, with certain assumptions, finding the most likely estimate of the posterior distribution of the explicit variable has an appealing intuitive interpretation (Freeman, 1994): *perception's model of the image should be robust over variations in generic variables* . This is a generalization of the generic view principle (Lowe, 1985; Nakayama and Shimojo, 1992), and follows from statistical decision theory. Note that as stated, a literal implementation would be top-down, because it would require measuring variations in the image domain. Section 2 shows how specifying depth and illumination direction as explicit and generic variables, respectively, reduces ambiguity in depth-from-shadows.

The posterior distribution defines the visual information available, but one still has to extract estimates and decisions according to some criteria. In Section 2, we estimate depth from shadows by picking the most probable value of the posterior probability. A visual task can also be a simple decision, which makes for good psychophysics. In Section 3, we ask humans and a statistically ideal observer to get the maximum percent correct answers to the question: "Is this right object?"

---

where the subscripts E and G indicate explicit and generic variables. With a loss function, $-\delta(\Sigma_E - S_E)$, where the cost to errors in the generic variable is constant, minimizing risk is equivalent to marginalizing the posterior with respect to the generic variable, and choosing the maximum of the posterior:

$$P(S_E \mid I) = \int P(S_G, S_E \mid I) dS_G$$

The importance of vision as Bayesian inference

The virtue of the Bayesian framework is that it requires one to describe all the assumptions which constrain the visual inference. Although a Bayesian analysis prescribes the constraints, it doesn't say how these should be embedded in visual mechanisms. It can, however, provide hints.

The rationale for using the Bayesian inversion of the posterior probability is that it is usually easier to specify the image rendering constraint, than the inverse visual inference problem. In other words, it is easier for the theoretician to write down the likelihood function which says how image information is determined from the scene, than the reverse. But does Bayes suggest more than a theoretical convenience? It has been argued that the inherent confounding of diverse scene causes in natural patterns, including images, necessitates analysis-by-synthesis through a generative model which tests top-down predictions of the input. One commonly discussed explanation for the pattern of back-projections between cortical areas is that these connections enable the expression of unresolved high-level hypotheses in the language of an earlier level (Mumford, 1994; Dayan, 1994). This expression can then be tested with respect to the incoming data at the earlier level. Thus, domain-specific models in memory can be manipulated to check for fits to the incoming data in ways that are difficult bottom-up. We return to this issue below in the discussion of human object recognition mechanisms.

Applying the statistical inference approach to high-level vision

We have seen that characterizing the statistical requirements for reliable visual inferences is a complex problem, because it requires modeling the signals the world is sending about object shape, material, and location, the way in which the signals get "muddled" in the form of an image, as well as the optimal means to decode this image. Solving these problems using Bayes methods for general purpose vision is not yet feasible. Practical applications to high-level human vision require: 1) a judicious approximation of a natural visual task and qualitative analysis of natural constraints; or 2) designing a computable psychophysical task, an approach with a successful

history in studies of early visual mechanisms (e.g. (Barlow, 1962; Geisler, 1989; Knill, in press; Pelli, 1990; Schrater, Knill, and Simoncelli, accepted pending revisions), and recent applications to the high-level visual tasks of reading (Legge, Klitz, and Tjan, 1997) and object recognition (Tjan et al., 1995).

Section 2 describes a qualitative analysis of depth-from-cast-shadows, a problem in spatial layout where resolving ambiguity in edge identity is particularly crucial (Knill, Kersten, and Mamassian, 1995). Realistic computer graphics allows the approximation of natural complexity, while retaining sufficient simplicity to analyze the image ambiguities and identify natural constraints. Section 3 describes an investigation of the problem of viewpoint variation in object recognition, where the second approach, often called *ideal observer* analysis, is adopted (Liu and Kersten, in press; Liu, Knill, and Kersten, 1995). A key issue is how recognition overcomes the geometrical problem of projection. Ideal observer analysis provides a rigorous means to normalize human performance with respect to the informational limits imposed by the task itself, and thereby draw firm conclusions about the underlying mechanisms.

2. *Depth-from-cast shadows: Qualitative analysis of constraints*

One can list well over a dozen cues to depth, including stereo disparity, motion parallax, and the pictorial cues. One of the pictorial cues, depth-from-cast-shadows is particularly interesting because it is surprisingly strong, and seems to involve a complex set of inferences (Kersten et al., 1996; Kersten, Mamassian, and Knill, 1997). To investigate depth-from-shadows, Kersten et al. (1996) made a movie of a square in front of a stationary checkerboard illuminated with an extended light source (Figure 4, top panel). The central square was fixed in the image, and the only movement was that of a shadow translating diagonally away and back to the square. Despite the lack of any image motion of the square, observers report an initial strong perception of the square moving in depth. The computer animation looks realistic and the perceptual interpretation unique, yet the image data have significant ambiguities of material and depth (Figure 5), and of light source motion and direction (Figure 6). This simple percept involves a range of decisions across several levels of abstraction. Let's consider in turn inferences of: context, motion event, image region categorization, and depth parameter estimation.

(1) A key question of context is : Which of the objects or viewer provides the frame of reference with which to interpret the locations of the other object(s)? Viewpoint should be generic for deciding whether the square is headed away from the checkerboard, but explicit if the task is to reach to the square. The central square, not the background, appears to move. Relative size, enclosure, and occlusion information in the image may all provide support for the decision that the checkerboard provides a stationary frame of reference. The decision that the background is opaque (see Figure 5) must involve a prior default on material, because the same image could have resulted from a transparent background and opaque "shadow"--a percept which can be seen given training (Kersten et al., 1992).

(2) At some level, an object-shadow "event" must be identified. This could involve combining independent identifications of surface and shadow image regions, or using global image information. Other experiments suggest that characteristic correlated motion, an image formation constraint, may be a global diagnostic for a moving object-shadow pair (Kersten, Mamassian, and Knill, 1997). For moving objects, the linkage between the object and its shadow is strongly constrained by a prior assumption that light sources usually don't move. The assumption that light sources are usually from above accounts for the finding that shadows above the object are less effective than those below (Figure 4)--an assumption well-known for shape-from-shading (Gibson, 1950).

(3) Computing relative depth depends on either an explicit or implicit categorization of image regions as: opaque surface, transparent surface or shadowed surface. A particularly diagnostic cue for motion in depth, is the changing fuzziness of the penumbra caused by an extended light source. This was the most effective condition in Figure 4. Such a local image measurement has less ambiguity with other scene causes (e.g. it is likely to be confused with a material change, although it could result from surface edge motion out of the depth-of-field range, or a spreading stain). This cue is also robust over viewpoint, and a large range of types of illumination. In contrast, the sharp shadow is often seen as a transparent surface--a decision also supported by local transparency constraints at X-junctions (Metelli, 1975). Physically unnatural light "shadows" violate local transparency constraints consistent with shadows, and lead to less effective apparent motion (Kersten, Mamassian, and Knill, 1997). Occlusion of the shadow by the object is a potentially important constraint for determining which patch of the object-shadow pair is the shadow; however, occlusion isn't necessary for depth-from-cast-shadows.

Local cues supporting a shadow hypothesis have to be weighed against the conflicting cues regarding motion in depth. Size change and velocity in the image are both zero, indicating no motion in depth. This is a consequence of viewpoint being a generic variable--the alternative

interpretation is of a square moving directly along the line of sight; but this is normally ruled out because small changes in viewpoint would produce large changes in the image. The fact that depth change with shadow motion is seen is evidence of a strong prior stationary light source constraint.

(4) A visual estimation can be made as to the square's location or velocity from the measured shadow location or velocity. Suppose the object-shadow pair is detected, the shadow identified and localized to the background surface. The stationary light source assumption would resolve ambiguity regarding motion in depth. But what about the stationary case? Where is the light source?

The visual task constrains an estimate of relative depth: robustness with respect to generic variables

Consider the simple geometric ambiguity illustrated in Figure 6. The measured displacement, x, between the image of an object and its shadow can be caused by an infinite number of combinations of object distance, z, and light source direction, α :

$$x = z\tan(\alpha)$$

An additional constraint is required to estimate z from x. One could try to measure (or make up) a prior on the light direction that would produce a unique estimate of z. But the task itself provides a sufficient constraint to uniquely estimate z. Assume that the explicit variable is relative depth z, and the generic variable is light source direction, α. By differentiating the above geometric constraint on object, shadow, and light source parameters, we have:

$$\Delta x = \Delta\alpha \frac{x^2 + z^2}{z}$$

For a given variation $\Delta\alpha$, the minimum change in Δx would occur for z=x. Perception's estimate of shadow displacement is most robust to variations in light direction for relative depths equal to the displacement, i.e. equivalent to assuming the light is at 45 degrees[2].

[2]Assuming no image noise and a uniform distribution for z and α., the mode of $p(z|x)$ is also z=x. Marginalize over the generic variable α:

$$p(z \mid x) = \int_{-\pi/2}^{\pi/2} p(z,\alpha \mid x) d\alpha \qquad \int_{-\pi/2}^{\pi/2} \delta(x - z\tan(\alpha)) d\alpha = \frac{z}{x^2 + z^2}, \qquad 0 \le z \le z_{\max}$$

3. Viewpoint compensation in 3D object recognition: Ideal observer analysis

A basic component of 3D object recognition is a process that verifies matches between the input stimulus and stored object representations in memory. The problem is that the images of a single object are enormously variable, depending on viewpoint, among other factors. The visual system must somehow compensate for such variations in order to identify an object as the same when seen from another viewpoint. There has been recent debate regarding the nature of these stored representations and the mechanisms which test for a match. On the one hand, certain object properties such as edge straightness are preserved in the image over viewpoint changes, suggesting that the early extraction of such features could be used fairly directly (Biederman, 1987; Hummel and Biederman, 1992). On the other hand, computer vision has shown the difficulties in extracting features such as edges from natural images. Further, the experimental observation that familiar views of an object are processed more effectively than unfamiliar views suggests that the memory of an object may be closely tied to images previously seen of that object (Bülthoff and Edelman, 1992; Tarr and Bülthoff, 1995).

How does the visual system compensate for image variations in size, position, and rotation in depth produced by an object? One can devise schemes to allow for variations in scale and position through feedforward mechanisms (Ullman, 1996). Neurons insensitive to object scale and position have been found in inferotemporal cortex of monkeys (Logothetis et al., 1994). Allowing for rotations in depth, however, seems more problematic because depth information is lost in the 2D projection, so one doesn't know how to transform the image to allow for these rotations. Let's consider two cases that differ in the degree to which 3D information is involved in a test for a match.

Suppose that an object is represented as a collection of independent 2D images or views in memory. These views have, through experience, come to be associated with each other, and have a

common label. In order to recognize a novel view, similarity is measured independently between this novel view and each of the familiar views. The combination of the measurements determines if the novel view should be recognized or rejected. Although the measure of similarity has some flexibility, the crucial point is that recognition can be achieved with 2D manipulations of the images without reconstructing the 3D structure of the object either explicitly or implicitly (Bülthoff and Edelman, 1992; Poggio and Edelman, 1990). Below we describe a smarter version of such a model for human vision (the 2D/2D observer), which in addition allows for possible rotations in 2D for each template view.

Contrast this with a second case in which there is a candidate 3D object model in memory. Then the appropriate transformation could be applied to the model in memory, and thus compensate for rotations in depth in order to test for a match. Imagine two sub-cases. The most straightforward identification scheme verifies a match by translating, scaling, and rotating an explicit 3D model of the object in memory, projecting the result in a 2D image space, and then using a measure of similarity to test for a satisfactory match with the 2D input (Basri and Weinshall, 1996). The statistically optimal version of this model is called a 3D/2D ideal observer (Liu, Knill, and Kersten, 1995). Despite its intuitive simplicity, a straightforward implementation is computationally unrealistic even for simple objects--the space of transformations is just too big. However, a clever shortcut was discovered by Ullman and Basri (1991)--with as few as two views one could carry out the verification process by checking the linear dependence of the input image on the two stored views.

Liu, Knill and Kersten (1995) devised a 3D object discrimination task for which they could calculate ideal performance for the 2D/2D and 3D/2D classes of observers. By comparing human with ideal performance, they were able factor out limitations imposed by the task itself, and thereby investigate how the human visual system compensates for viewpoint change.

Their object world was simple: Five randomly placed vertices (3D points) were connected by four straight cylinders of uniform diameter, making 3D wire prototype objects that looked like bent paper-clips. A pair of objects was generated from a prototype by adding independent 3D

positional Gaussian noise at the vertex points. One object is called the target, whose Gaussian noise has a fixed variance. The other is called the distractor, whose variance is always larger. In the test phase (see below), the novel views of an object could have any orientation in space--i.e., the prior distribution on rotations in 3D was uniform. The 3D/2D observer has complete knowledge of the target object and task. Prior knowledge of the target object is given in the form of 11 views to the 2D/2D observer, and to the human observers through training.

In contrast to the analysis of depth-from-cast-shadows, the task requirements for the ideal and human observers are precisely specified. Optimal recognition performance is based on the shape of the object defined by the image vertex positions (explicit variables), with viewpoint variables as generic. The task is summarized in Figure 7. Both the human and ideal observers must choose from the two images an object that is more similar (in Euclidean distance) of the feature points to the prototype object.

3D/2D ideal observer

Let's formalize the inference constraints for the 3D/2D ideal observer. Occlusion can be neglected because the vertex feature points for wire objects are visible from almost all viewing angles. Further, because the vertices are connected, one knows how to order the vertices when comparing stimulus to memory. The visual decision is based on a representation of the objects and images in terms of 15 and 10 dimensional vector vertex locations, respectively.

The 3D/2D ideal observer matches the stimulus image against all possible views of a known prototype object. By definition, the ideal's image rendering model is:

$$\mathbf{I} = F_{\Phi}(\mathbf{O}) + \mathbf{N}_p$$

where $\mathbf{I}$ and $\mathbf{O}$ are representations of the 2D vertex positions of the 2D image, and 3D vertex positions of the object, respectively. $F_{\Phi}(\cdot)$ represents the combined effects of an unknown viewpoint transformation in 3D (represented by a three-component vector Φ), followed by orthographic projection. $\mathbf{N}_p$ is the positional noise of the projected vertex positions. An ideal

observer which can only detect the 2D vertex positions in a stimulus image, but has a full 3D model of the prototype would estimate the probability of obtaining image $\mathbf{I}_k$ from the target (smaller noise variance) by integrating out the generic viewpoint variables to obtain the probability of $\mathbf{I}_k$:

$$p_t(\mathbf{I}_k) = \int p(\mathbf{N}_p = \mathbf{I} - F_\Phi(\mathbf{O}))p(\Phi)d\Phi$$

To achieve the maximum average percent correct, the ideal observer chooses the image (k=1 or 2) with the bigger value of $p_t(\mathbf{I}_k)$.

The essence of the 3D/2D ideal observer is that it has an exact model of the 3D object, **O**, in memory, as well as precise knowledge of how such an object in the world could be transformed into an image, **I**. This transformation includes the unknown generic variables of rotation. A key component in the ideal calculation is a measurement of similarity, which because the noise is Gaussian is given by: $\|\mathbf{I} - F_\Phi(\mathbf{T}_i)\|^2$. In theory, a straightforward implementation of the probability calculation would involve manipulations in a 3D object space followed by back-projection of the model into image space to measure the similarity. (Again, because of the large transformation space, this calculation isn't feasible; see Liu, Knill, and Kersten, 1995 for an approximation). The 2D/2D observer (below) is an alternative way of measuring similarity which relies on manipulations that can, in principle, be done entirely in a 2D image space. It's efficiency is less than the 3D/2D observer's, but can it account for human performance?

The 2D/2D observer

In the experimental task, an observer sees 11 distinct views of the object--familiar views from which a 3D/2D observer could in theory construct its 3D object model, **O** (11 is more than enough to do this). Suppose, however, that there was no mechanism to construct such a model, and the recognition system had to rely on making matches of the 11 familiar views in memory to the stimulus image. Further, suppose it had available rigid rotations in the 2D plane to compensate as best it could for the normal image variations that arise through 3D rotations. This 2D/2D

observer has the wrong image rendering model. Yet it does its best by optimally combining information from stored multiple views under the constraint of being limited to 2D rigid transformations specified by a rotation matrix, R_φ.

Let **I** represent the coordinates of the vertices in a stimulus image, and **T** = {$\mathbf{T}_1$, $\mathbf{T}_2$, …, $\mathbf{T}_{11}$} represent the 11 prototype templates. Now $p_t(\mathbf{I})$ is given by:

$$p_t(\mathbf{I}) = \sum_{i=1}^{11} \int_0^{2\pi} [p(\mathbf{I} - R_\varphi(\mathbf{T}_i) p(R_\varphi(\mathbf{T}_i)) d\varphi$$

where $p\ (\mathbf{I} \mid R_\varphi(\mathbf{T}_i))$ is the probability that **I** was generated by adding noise to template $\mathbf{T}_i$ at 2D rotation angle φ. The prior probability, $p(R_\varphi(\mathbf{T}_i))$ is constant (=1/2π) [3]. Because a rotation of the model ($R_\varphi(\mathbf{T}_i)$) is equivalent to an inverse rotation of the image ($R_\varphi^{-1}(\mathbf{I})$) for the 2D/2D observer, variation over viewpoint can be compensated for either by feedforward, or back-projection.

Human performance

Figure 8 shows human performance relative to the 3D/2D and 2D/2D observers, for decisions based on both familiar and novel views. Performance is measured in terms of statistical efficiency[4]. The 3D/2D efficiency factors out the limits to performance imposed by the task itself, independent of any algorithm used to compute decisions. The 2D/2D efficiencies are expressed in the same units.

First note that 3D/2D efficiencies are not 100%. There are two main sources of inefficiency for humans: intrinsic noise, and an inappropriate transformation process. If the only problem was some internal uncertainty added to the artificially introduced positional noise, then the efficiencies

---

3 The actual calculation was slightly more complicated to allow for uncertainty as to which vertex was first.

4 Statistical efficiency is defined as the ratio of the number of data samples the ideal requires to the number the human observer requires for an identical level of performance (e.g. same percent correct; see Liu, Knill, and Kersten, 1995).

for both familiar and novel views would be the same. The fact that novel views are dealt with less efficiently is consistent with theories of recognition that assume the visual memory for an object is closely tied to its stored familiar views.

The 2D/2D observer is a precise definition of one such view-dependent model. But now note that the statistical efficiencies for novel views are too high. In fact, efficiencies over 100% means that any implementation which verifies matches using remembered templates and rigid image manipulations can be excluded as a model for human performance.

These results show that human recognition uses a much “dumber” view-compensation mechanism than the 3D/2D ideal observer, but a “smarter” one than an independent comparison with stored views. One candidate smarter model would be to allow for 2D affine transformations that include translation, rotation, scale, and skew adjustments in the image domain. Recent work suggests that even this kind of view compensation is not sufficient to account for performance in this experiment (Liu and Kersten, in press).

4. Conclusions

We have seen how two quite different high-level visual functions, perception of depth and object recognition, can be investigated within the common framework of statistical inference. The perception of depth-from-cast-shadows involves a remarkable synthesis of default prior assumptions on material and lighting, local image cues and global constraints. We've considered just one of over a dozen sources of information for depth. A quantitative computational model for the perception of spatial layout is an important challenge for future vision research.

A key problem in 3D object recognition is understanding how the brain compensates for variations in viewpoint. By designing a relatively simple visual task for which the optimal inference is computable, one can pit human and ideal observers against each other in the same task. While the computational formulations can be demanding, ideal observer analysis has the potential to rigorously test well-defined models of human high-level functions. Statistical efficiency normalizes performance with respect to the information in the task, with remaining differences diagnostic of processing mechanisms of the visual brain. This research has shown that independent comparisons of images to templates in memory cannot account for human viewpoint compensation, even with some flexibility (via 2D rigid transformations) allowed in the matching process.

ACKNOWLEDGMENTS

The author's research is supported by the National Science Foundation (SBR-9631682) and the National Institutes of Health (RO1 EY11507-001). I also thank Zili Liu, Cindee Madison, Paul Schrater, and Brian Stankiewicz for their help.

FIGURE LEGENDS

Figure 1. Constraints on visual inference. Information about object shape, articulation, material, illumination, and viewpoint is encrypted in the image through rendering and projection. Diverse visual tasks depend on estimates of these scene variables. Some scene variable estimates are more important for some tasks than others. The important variables for a task are the *explicit* variables. Variables to be discounted are referred to as *generic*. For example, it is commonly assumed that shape, but not viewpoint, illumination or material, should be estimated explicitly for basic-level recognition (i.e. deciding whether an image is that of a dog, rather than a particular dog, “Snuggles”). Viewpoint and illumination are generic variables for object recognition at all levels. Hypothesized explicit variables are indicated in parentheses for various visual tasks (top box). The visual task, the nature of the projection of the scene onto the image, and the scene structure probabilities characterize the knowledge required for decoding image data.

Figure 2. Local contour constraints determine a global percept. Consider the upper left panel. The picture is usually seen as the overlapping profiles of two simians. This interpretation depends on how the four lines meeting at the two "X-junctions" are grouped. A local constraint of colinearity groups the X-junction into two crossing straight lines which is consistent with the simian percept. If the two halves are separated down the middle (lower left panel), one can easily see the other interpretation of two homo sapiens. Local constraints on transparency also affect how the contours are grouped (two right panels).

Figure 3. A measurement of a local change of image intensity, illustrated by the elliptical patch in the upper left, is highly ambiguous as to what in the scene caused it. A change in material, depth, surface orientation, specularity, or shadow can create the same local oriented intensity change, up to a spatial scale factor. (Adapted from Kersten, 1997).

Figure 4. Depth-from-cast-shadows. Observers viewed computer animations in which a central square was held fixed in the image, while its shadow moved diagonally back and forth. The simulations were produced by moving the central square back and forth directly along the line of sight; thus, under orthographic projection, the image of the square does not change size or move. The upper panel shows first and last frames for the main condition, in which the illumination was from an extended light source above the square. The middle panel shows final frames for this condition, including three others. From left to right, the conditions are: extended light source from above, extended light source from below, point light source from above, and point light source from below. The extended light source (like a fluorescent panel) produces a penumbra that gets fuzzier as the square gets further away from the background. Despite the lack of objective image motion of the central square, it nevertheless almost always appeared to move in depth for the extended light from above condition. The bar graph (lower panel) shows the proportion of times (out of 15) observers reported the central square patch to be moving in depth for extended and point light sources from above or below. There is a significant advantage of an extended light vs. a point source ($z=2.28$, $p<.02$), and of light from above vs. below ($z=3.028$, $p<.002$). A QuickTime™ movie demonstrating illusory motion from shadows can be viewed and downloaded from: http://vision.psych.umn.edu/www/kersten-lab/shadows.html. (Adapted from Kersten et al., 1996).

Figure 5. Ambiguities of material and spatial layout for the depth-from-shadows movies. Assume the central square, the “shadow”, and background of Figure 4 have been segmented, but not labeled according to whether they are opaque material, transparent material, or shadow. Image formation constraints guarantee that the central square and the dark “shadow” regions lie somewhere along the line of sight--but where? The background could be transparent and in front of

the “shadow” (**a**), rather than the reverse (**b** or **c**). If outside the eye's depth-of-field, the ersatz shadow image would mimic the fuzziness of a penumbra change. If the “shadow patch” was instead a transparent surface, it could be at location **b** or **c**; but if a shadow, it would have to be at **b**. Occlusion cues place the square in front of the background. For a reliable inference of depth-from-shadows, the shadow has to be labeled as such, localized to the background, and linked to the casting object. And then one is still left with the question of where the light source is, and if motion in the image is due to movement of the light source or the central square (see Figure 6).

Figure 6. The visual task constrains the estimate of relative depth-from-shadows. This figure shows a very simplified view of the geometrical constraints where we’ve assumed that the remaining unknowns are the light source direction, α and relative depth, z. By treating light source direction as a generic variable, one can show that the best bet for the target depth is $z = x$.

Figure 7. The 3D object classification task presented to human and ideal observers. The observers are required to discriminate between two classes of wire object - one generated by adding a small fixed amount of noise to the vertices a prototype object (the target), and the other generated by adding a larger amount of noise to the prototype object (distractor). The means for both target and distractor sets are the same prototype object. Two stimuli were generated by a 3D rotation of the noiseless prototype. The standard deviation of the positional noise added to the distractor (prototype + more noise) was greater than that added to the signal (prototype + noise). Knowledge of the wire objects was provided in a prior training session in which the object prototype was first learned from a discrete number (11 rotations) of its images. The 11 training views of a prototype object were created by rotating the object first around the X-axis (horizontal in the screen plane) six times in 60 degree steps, and then around the Y-axis (vertical in the screen plane) six times, again with 60 degree rotational steps, resulting in 11 views of the object. The angle with the Z-axis was chosen from a uniform distribution between 0 and 180 degrees and the angle with X-axis was chosen from a uniform distribution between 0 and 360 degrees.

Figure 8. Statistical efficiencies for human performance relative to the 3D/2D and 2D/2D observers. The means were computed by averaging the efficiencies across the three objects for each type. The error bars show ± one standard deviation.

REFERENCES

Barlow, H. B., 1962. A method of determining the overall quantum efficiency of visual discriminations. J. Physiol. (Lond.) 160: 155-168.

Basri, R. and D. Weinshall, 1996. Distance metric between 3D models and 2D images for recognition and classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 18: 465-470.

Belhumeur, P. N., J. P. Hespanha, and D. J. Kriegman, 1996. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. In European Conference on Computer Vision.

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

Blake, A. and A. Yuille, 1992. Active Vision. Cambridge, MA: MIT Press.

Brainard, D. H. and W. T. Freeman, 1994. Bayesian Method for Recovering Surface and Iluminant Properties from Photosensor Responses. In Human Vision, Visual Processing, and Digital Display V, 2179:364-376. Bellingham, Washington: The Society of Photo-Optical Instrumentation Engineers.

Bülthoff, H. H. and S. Edelman, 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc. Natl. Acad. Sci. USA 89: 60-64.

Clark, J. J. and A. L. Yuille, 1990. Data Fusion for Sensory Information Processing. Boston: Kluwer Academic Publishers.

Dayan, P., G. E. Hinton, R. M. Neal, and R. S. Zemel, 1995. The Helmholtz Machine. Neural Computation 7 (5): 889-904.

Freeman, W. T., 1994. The generic viewpoint assumption in a framework for visual perception. Nature 368 (7 April 1994): 542-545.

Geisler, W., 1989. Sequential Ideal-Observer analysis of visual discriminations. Psychological Review 96 (2): 267-314.

Gibson, J. J., 1950. The Perception of the Visual World. Boston, MA: Houghton Mifflin.

Hummel, J. E. and I. Biederman, 1992. Dynamic binding in a neural network for shape recognition. Psychological Review 99 (3): 480-517.

Kersten, D., 1990. Statistical limits to image understanding. In Vision: Coding and Efficiency, ed. C. Blakemore:32-44. Cambridge, UK: Cambridge University Press.

Kersten, D., H. H. Bülthoff, B. Schwartz, and K. Kurtz, 1992. Interaction between transparency and structure from motion. Neural Computation 4 (4): 573-589.

Kersten, D., D. C. Knill, P. Mamassian, and I. Bülthoff, 1996. Illusory motion from shadows. Nature 379: 31.

Kersten, D., P. Mamassian, and D. C. Knill, 1997. Moving cast shadows induce apparent motion in depth. Perception 26 (2): 171-192.

Kersten, D. J., 1991. Transparency and the cooperative computation of scene attributes. In Computational Models of Visual Processing, ed. M. Landy and A. Movshon:209-228. Cambridge, Massachusetts: M.I.T. Press.

Kersten, D., 1997. Inverse 3D Graphics: A metaphor for visual perception. Behavior Research Methods, Instruments, & Computers 29 (1): 37-46.

Knill, D. C., in press. Surface orientation from texture: Ideal observers, generic observers and the information content of texture cues. Vision Research .

Knill, D. C., D. Kersten, and P. Mamassian, 1995. The Bayesian framework for visual information processing: implications for psychophysics. In Perception as Bayesian Inference, ed. K. D.C. and R. W.:Chap. 5: Cambridge University Press.

Knill, D. C. and W. Richards, 1996. Perception as Bayesian Inference. Edited by D. C. Knill and W. Richards. Cambridge: Cambridge University Press.

Landy, M. S., L. T. Maloney, E. B. Johnston, and M. J. Young, 1995. Measurement and modeling of depth cue combination: In defense of weak fusion. Vision Research 35: 389-412.

Legge, G. E., T. S. Klitz, and B. S. Tjan, 1997. Mr. Chips: an ideal-observer model of reading. Psych. Review 104 (3): 524-53.

Liu, Z. and D. Kersten, in press. 2D observers for 3D object recognition? Vision Research .

Liu, Z., D. C. Knill, and D. Kersten, 1995. Object classification for human and ideal observers. Vision Research 35 (4): 549-568.

Logothetis, N. K., J. Pauls, H. H. Bülthoff, and T. Poggio, 1994. View-dependent object recognition in monkeys. Current Biology 4 (5): 401-414.

Lowe, D. G., 1985. Perceptual Organization and Visual Recognition. Kluwer International Series in Engineering and Computer Science. Robotics : Vision, Manipulation: Kluwer Academic.

Metelli, F., 1975. Shadows without penumbra. In Gestaltentheorie in der modernen psychologie, ed. S. Ertel, L. Kemmler, and L. Stadler:200-209. Darmstadt: Dietrich Steinkopff.

Mumford, D., 1994. Neuronal architectures for pattern-theoretic problems. In Large-Scale Neuronal Theories of the Brain, ed. C. Koch and J. L. Davis:125-152. Cambridge, MA: MIT Press.

Mumford, D., 1995. Pattern theory: A unifying perspective. In Perception as Bayesian Inference, ed. D. C. Knill and R. W.:Chapter 2. Cambridge: Cambridge University Press.

Nakayama, K. and S. Shimojo, 1992. Experiencing and perceiving visual surfaces. Science 257: 1357-1363.

Olshausen, B. A. and D. J. Field, 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 607-609.

Pelli, D. G., 1990. The quantum efficiency of vision. In Vision:Coding and Efficiency, ed. C. Blakemore. Cambridge: Cambridge University Press.

Poggio, T. and S. Edelman, 1990. A network that learns to recognize three-dimensional objects. Nature 343: 263-266.

Poggio, T., V. Torre, and C. Koch, 1985. Computational vision and regularization theory. Nature 317: 314-319.

Schrater, P. R., D. C. Knill, and E. P. Simoncelli, accepted pending revisions. Mechanisms of visual motion detection. Nature .

Sha'ashua, A. and S. Ullman, 1988. Structural saliency: The detection of globally salient structures using a locally connected network. In 2nd International Conference on Computer Vision, 88:321-327. Washington, D.C.: IEEE Computer Society Press.

Tarr, M. J. and H. H. Bülthoff, 1995. Is human object recognition better described by geon-structural-descriptions or by multiple-views? Journal of Experimental Psychology: Human Perception and Performance 21 (6): 1494-1505.

Tjan, B., W. Braje, G. E. Legge, and D. Kersten, 1995. Human efficiency for recognizing 3-D objects in luminance noise. Vision Research 35 (21): 3053-3069.

Troje, N. F. and T. Vetter, 1996. Representations of human faces: Max Planck Institute for Biological Cybernetics. Technical Report No 041
ftp://ftp.mpik-tueb.mpg.de/pub/mpi-memos/TR-041.ps.Z

Ullman, S., 1996. High-level Vision: Object Recognition and Visual Cognition. Cambridge, Massachusetts: MIT Press.

Ullman, S. and R. Basri, 1991. Recognition by linear combinations of models. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (10): 992-1006.

Yuille, A. L. and H. H. Bülthoff, 1996. Bayesian decision theory and psychophysics. In Perception as Bayesian Inference, ed. K. D.C. and R. W. Cambridge, U.K.: Cambridge University Press.

Zhu, S. C., Y. Wu, and D. Mumford, 1997. Minimax Entropy Principle and its applications to texture modeling. Neural Computation 9 (8): 1627-1660.

Zucker, S. W. and C. David, 1988. The organization of curve detection: Coarse tangent fields and fine spline coverings. In Proceedings 2nd International Conference on Computer Vision. Tarpon Springs, Florida.

FIGURES

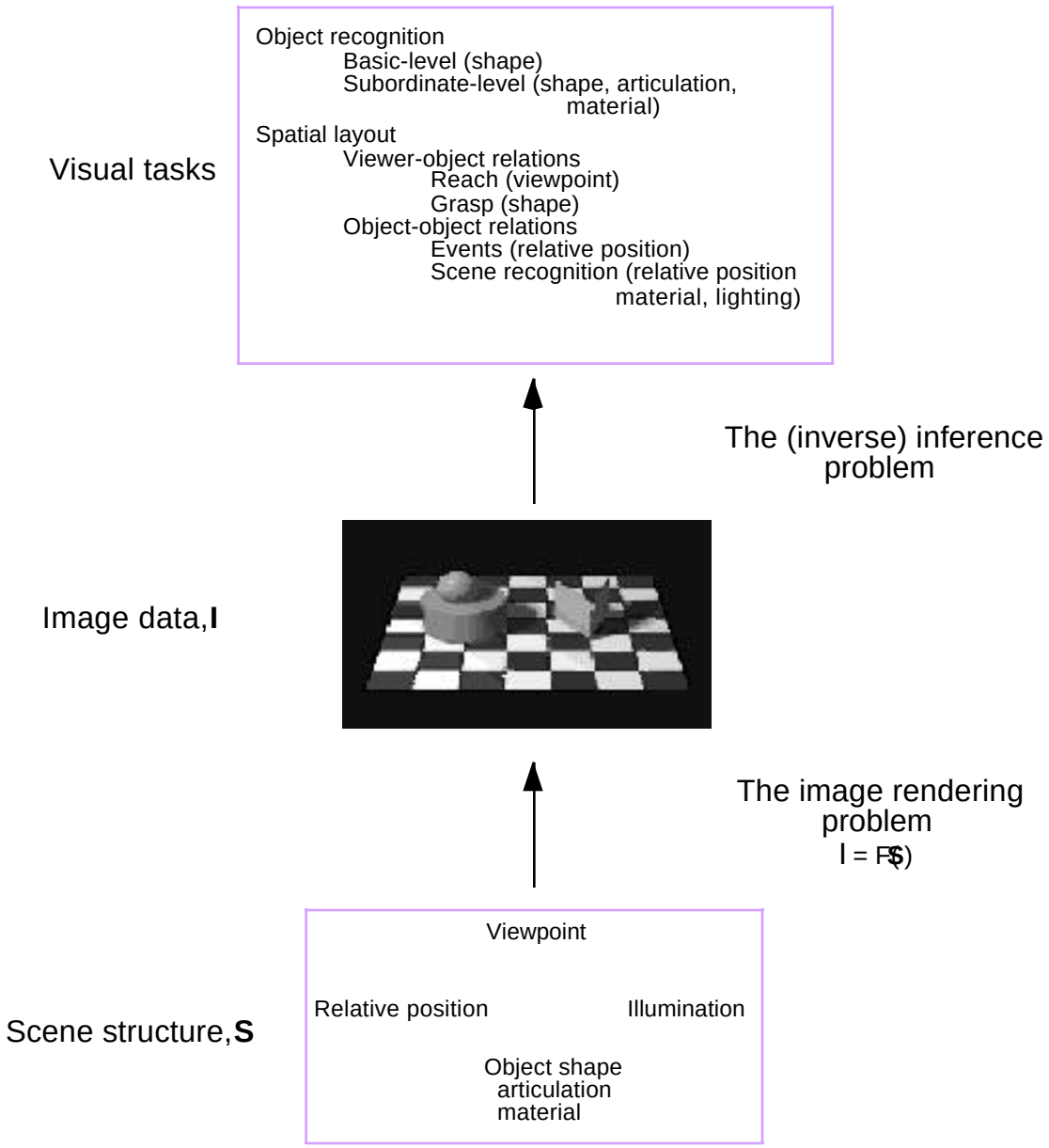

Figure 1

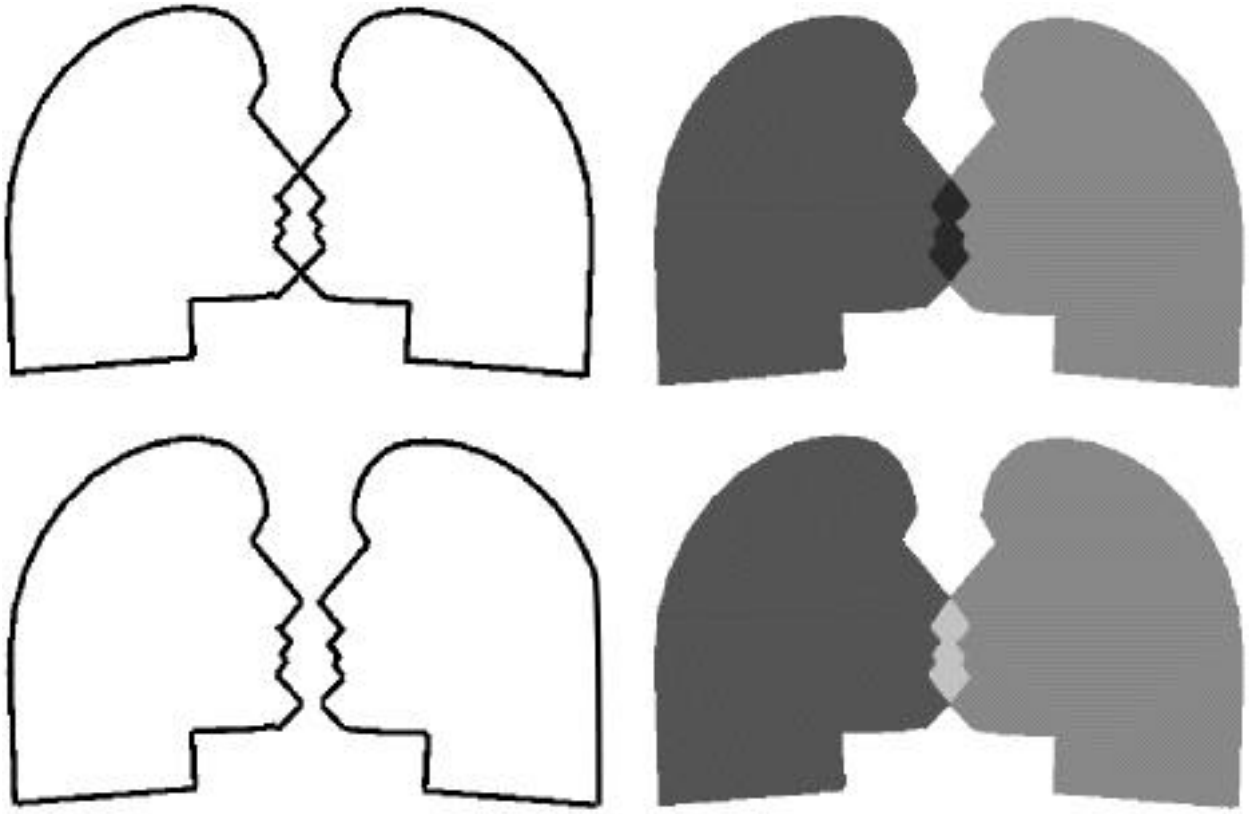

Figure 2

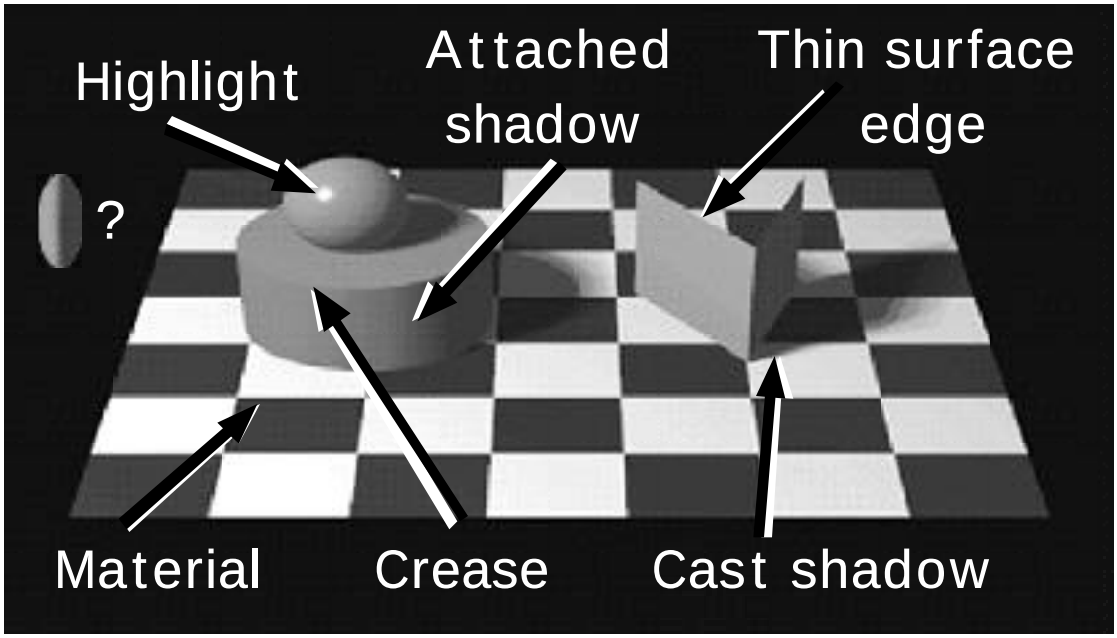
Highlight
Attached shadow
Thin surface edge
?
Material
Crease
Cast shadow

Figure 3

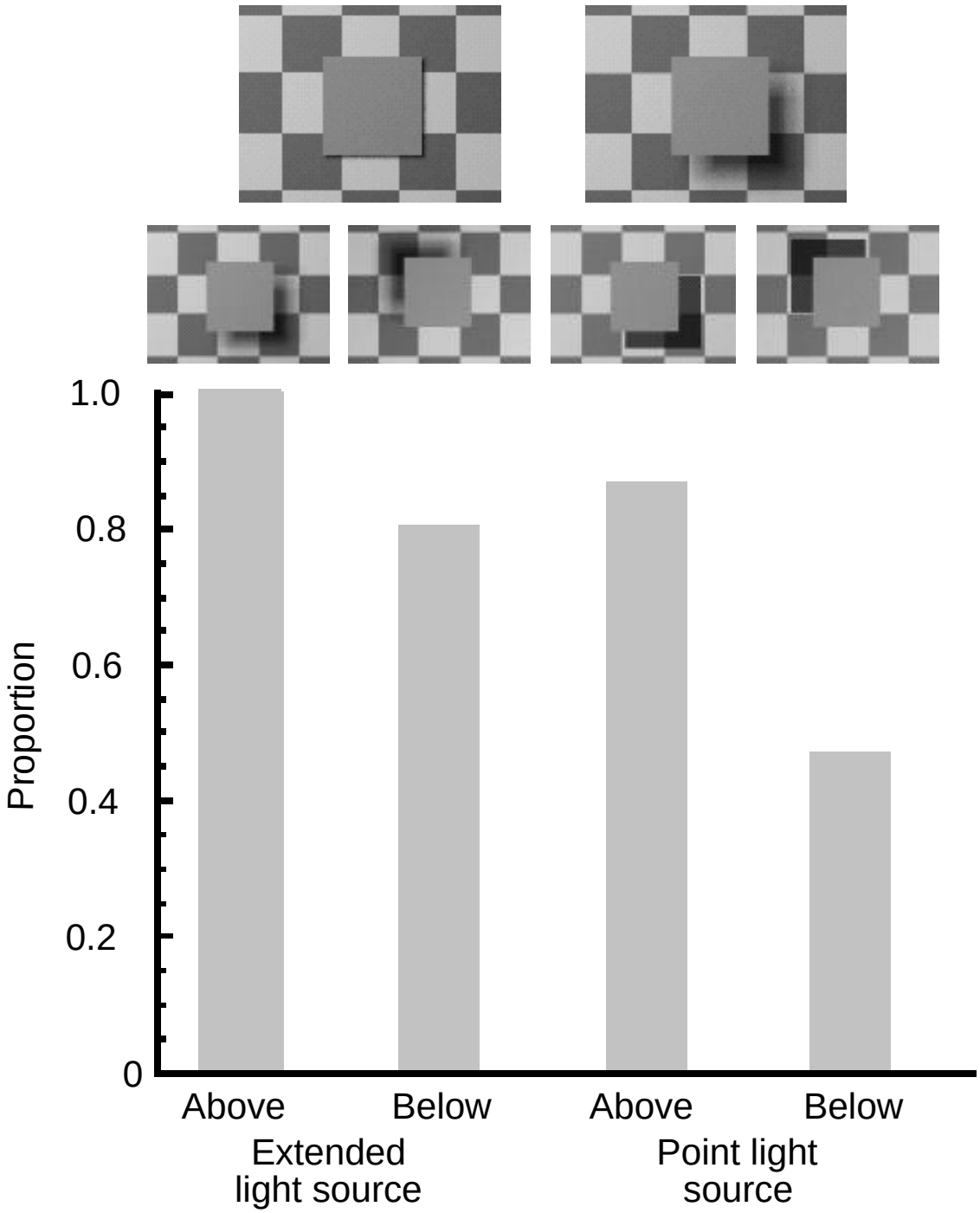
1.0
0.8
0.6
0.4
0.2
0
Proportion
Above
Below
Above
Below
Extended
light source
Point light
source

Figure 4

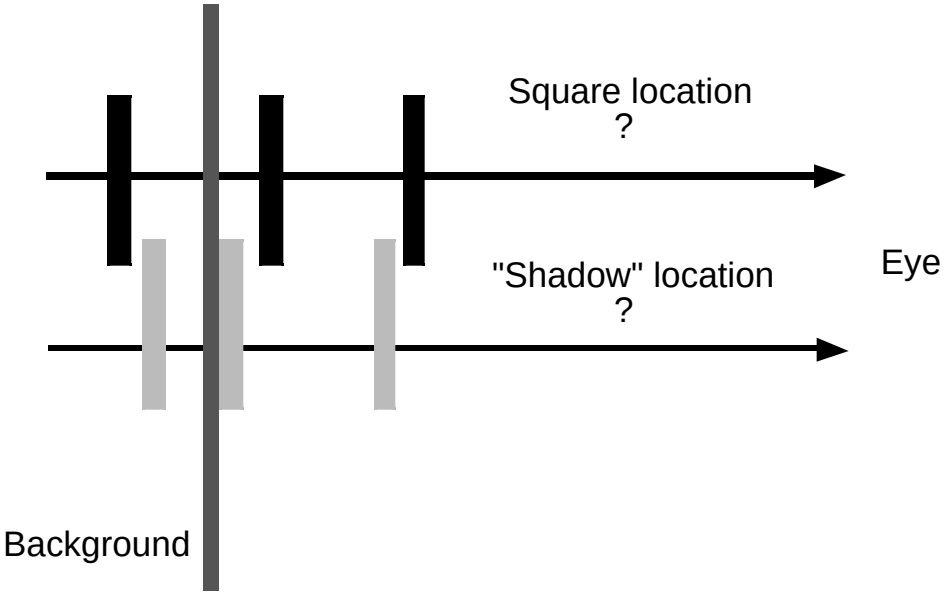

Figure 5

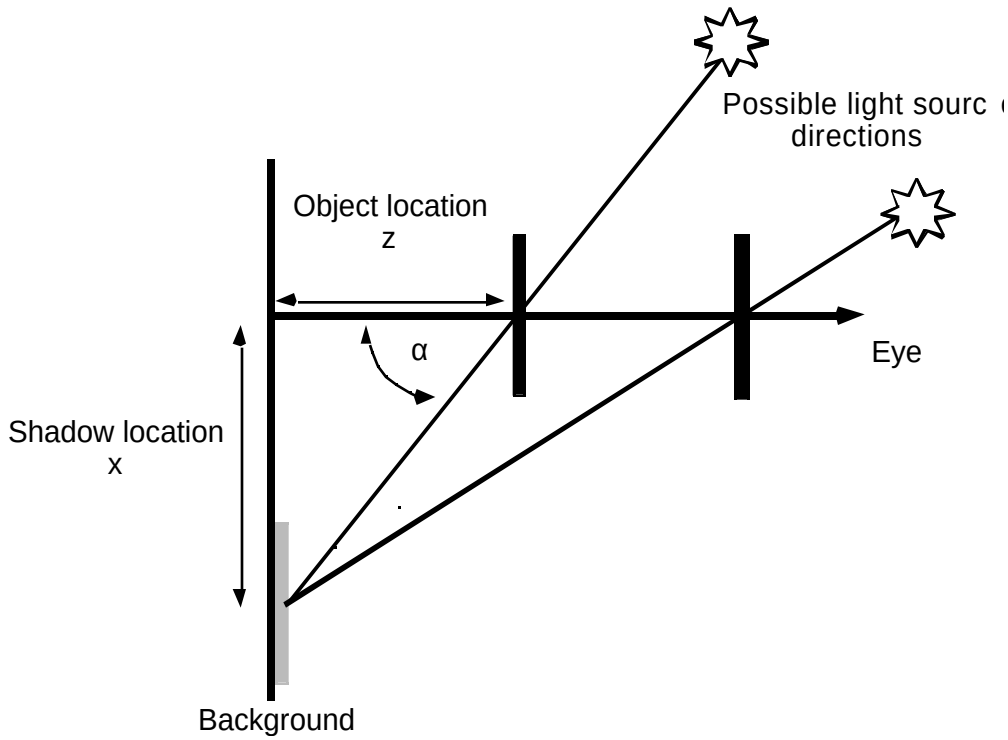
Possible light sourc
directions
Object location
z
α
Eye
Shadow location
x
Background

Figure 6

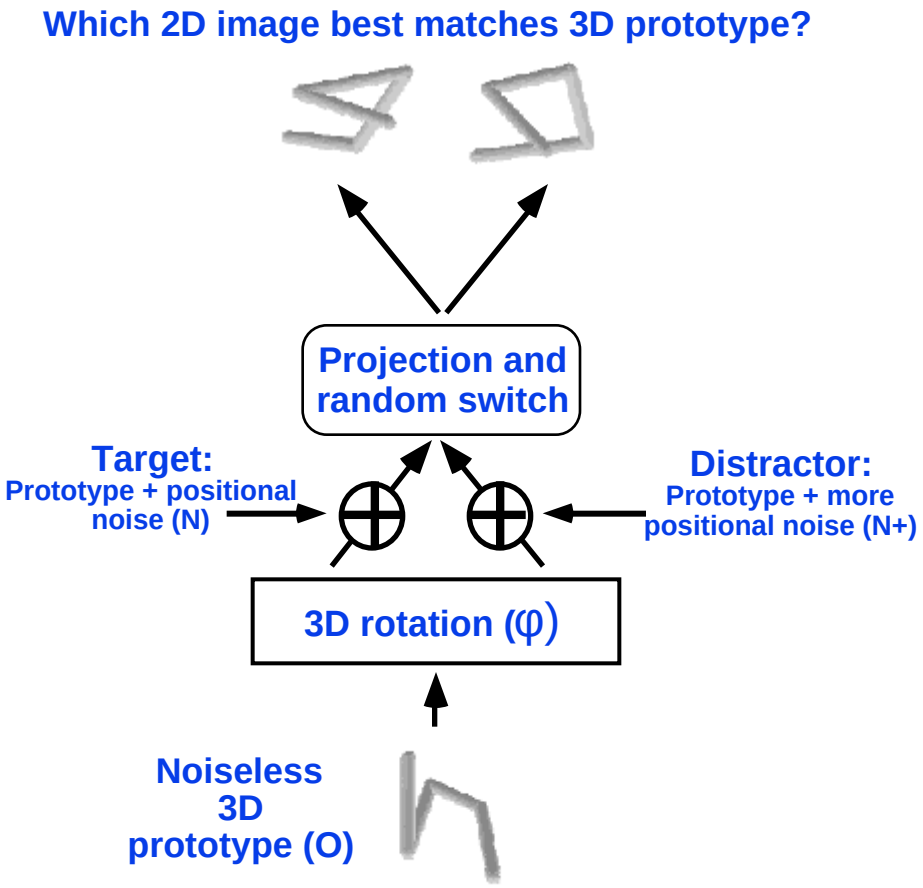
Which 2D image best matches 3D prototype?
Projection and random switch
Target:
Prototype + positional noise (N)
Distractor:
Prototype + more positional noise (N+)
3D rotation (φ)
Noiseless 3D prototype (O)

Figure 7

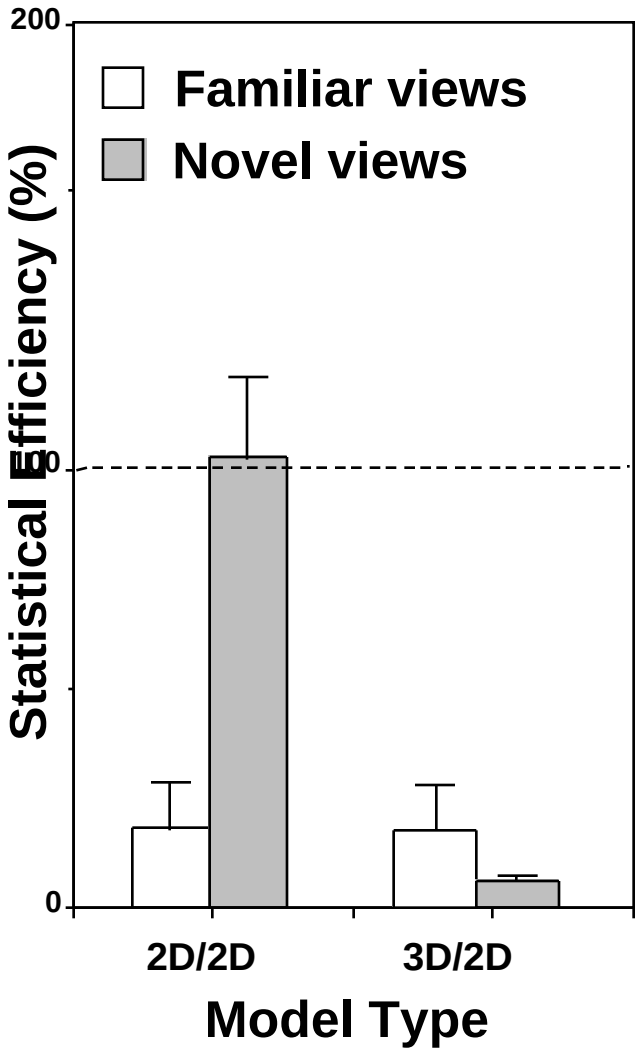
200
100
0
Statistical Efficiency (%)
Familiar views
Novel views
2D/2D
3D/2D
Model Type

Figure 8