Cognitive Science 28 (2004) 227–239

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Classification objects, ideal observers & generative models

Cheryl Olman[a,*], Daniel Kersten[b]

[a] *Graduate Program in Neuroscience, University of Minnesota, USA*
[b] *Department of Psychology, University of Minnesota, USA*

Received 31 January 2003; received in revised form 9 September 2003; accepted 25 September 2003

Abstract

A successful vision system must solve the problem of deriving geometrical information about three-dimensional objects from two-dimensional photometric input. The human visual system solves this problem with remarkable efficiency, and one challenge in vision research is to understand how neural representations of objects are formed and what visual information is used to form these representations. Ideal observer analysis has demonstrated the advantages of studying vision from the perspective of explicit generative models and a specified visual task, which divides the causes of image variations into the separate categories of signal and noise. Classification image techniques estimate the visual information used in a task from the properties of "noise" images that interact most strongly with the task. Both ideal observer analysis and classification image techniques rely on the assumption of a generative model. We show here how the ability of the classification image approach to understand how an observer uses visual information can be improved by matching the type and dimensionality of the model to that of the neural representation or internal template being studied. Because image variation in real world object tasks can arise from both geometrical shape and photometric (illumination or material) changes, a realistic image generation process should model geometry as well as intensity. A simple example is used to demonstrate what we refer to as a "classification object" approach to studying three-dimensional object representations.

Keywords: Classification image; Ideal observer; Reverse correlation

1. Introduction

The human visual system is confronted with a daunting task: make sense of data from the retina, in which a cluttered three-dimensional world is collapsed into a two-dimensional array

[*] Corresponding author. Tel.: +1-612-625-5372; fax: +1-612-626-2004.
E-mail addresses: cheryl@cmrr.umn.edu (C. Olman), kersten@umn.edu (D. Kersten).

0364-0213/$ – see front matter © 2003 Cognitive Science Society, Inc. All rights reserved.
doi:10.1016/j.cogsci.2003.09.004

of image intensities. A major goal of vision research is to understand how the visual system segments images into meaningful parts and produces a neural representation of the relevant aspects of three-dimensional objects. If images are treated just as a collection of pixels, then the dimensionality of image space—the number of ways in which an image can vary—is enormous (Kersten, 1987), and arriving at a succinct description is difficult, if not impossible. One way to restrict the dimensionality of image space is to define a generative model, which models images in terms of their sources—real world object arrangements and illumination conditions (Grenander, 1996; Kersten, 2002). The result is that the dimensionality of the parameter space for the generative model is much smaller than the dimensionality of the images it describes.

How can we use generative models to study the built-in assumptions that the human visual system has for its various visual tasks? Ideal observer and reverse correlation analysis, both originally developed in the context of signal detection theory, provide two tools for testing theories of information utilization in vision. We discuss how the notion of a generative model can be used to link and compare both techniques for studying human object perception, and how selecting generative models that are non-linear in image intensity space, but linear in three-dimensional object space, can extend classification image experiments to better probe information utilization in object recognition tasks.

To add a concrete example to the discussion, we will consider the following scene: a giraffe standing in front of a cluttered background of trees and bushes. The visual task is to decide what kind of an animal is in view; the vision researcher is interested in learning what visual information the observer uses to solve the task. Broadly speaking, visual information has two parts: the image measurements (or features), and the target (or signal) of interest. Image information about the signal, such as 3D description of a giraffe, is typically confounded by uncertainties (or noise) introduced by rendering and projection. One key aspect of this study is that the visual task defines which image variables are signal and which are noise. In this particular categorization task, image details such as the precise shapes of the spots on the giraffe or the leaves on the bushes are unimportant; what matter are features such as neck length and body size. However, if the task were identification of either the particular giraffe or the kind of bush, then the shapes of the spots and leaves would become important.

This example also offers an opportunity to clarify use of the term “noise,” which carries several different meanings. A recognition or segmentation task divides image attributes into the categories of explicit (signals, or relevant information) and generic (confounding, or irrelevant) variables. We will use the term *noise* to refer to variations in generic or confounding variables, the term *random deformations* to refer to variations in explicit variables, and *noise patterns* to refer to the specific noise patterns used in classification image (reverse correlation) experiments. While not discussed in detail in this paper, the term noise can describe not only the external variations but also to internal variations within the subject (Richards & Zhu, 1994).

1.1. Ideal observers: Bayesian inference of object properties

Ideal observer analysis measures the efficiency with which an observer uses information for a task by comparing the observer’s performance to the statistically best performance for the same task. The “observer” can be a neuron, or a human observer (Geisler & Albrecht, 1995; Watson, Barlow, & Robson, 1983). In our context, the ideal observer makes Bayesian optimal inferences about object properties from noisy image input, where noise is any cause of image variation that confounds the task-relevant object properties with task-irrelevant ones.

More formally, there are three key concepts in Bayesian inference. First, a generative model, $f : W \rightarrow I$, specifies how image features I are determined by the scene W, i.e., the object properties, relationships, illumination, and viewpoint. Second, the visual task splits W into components (S, N) that specify which scene properties are important to estimate explicitly (S, e.g., object shape), and which confound the measurements (N, e.g., viewpoint). Thus, $I = f(S, N)$. Third, the ideal observer is an inverse solution, $I \rightarrow S$, to the generative model that integrates diagnostic features, and discounts the confounding variables. Quantitatively, the generative model is specified by the likelihood of the image features, $P(I|S, N)$,[1] and the prior probability of the scene description, $P(S, N)$. The generative components determine a posterior probability distribution, $P(S, N|I)$ which is given by Bayes' rule:

$$P(S, N|I) = \frac{P(S, N)P(I|S, N)}{P(I)} \propto P(S, N)P(I - f(S, N))$$

where we assume additive noise, and $P(I)$ is fixed for a given image. Ideal observer decisions are based on the posterior $P(S|I)$, where N is discounted through integration over N, or if discrete, summing:

$$P(S|I) = \sum_{N} P(S, N|I)$$

This posterior probability defines the visual information available, and the ideal observer extracts estimates and makes decisions according to an optimality criterion, such as maximizing the proportion of correct decisions.[2] From the standpoint of inference, knowledge of prior constraints eliminates alternative image interpretations that might be consistent with the image data, but are unlikely, based on prior probability, $P(S)$. Note that the posterior characterizes the information, but does not prescribe a mechanism. Processes of inverse inference do not necessarily require a component that implements the generative model; nevertheless, some types of Bayesian inference (e.g., to deal with ambiguities of articulation or occlusion, cf. Ullman, 1996) may require an *internal* generative process (Grenander, 1996) that in some sense mirrors aspects of the external generative model in order to validate or rule out competing hypotheses.

Both components of the generative model, $P(S, N)$ and $P(I|S, N)$, can be modeled from physical intuition or learned from real-world statistics. Probability models relevant for object perception can be either geometric or photometric. They can be generic, such as for surface smoothness, shape, contour, or material (Poggio, Torre, & Koch, 1985), or be estimated from a specific object domain (Troje, 2002; Vetter & Poggio, 1997). Image (likelihood) models can be learned for specialized tasks, such as edge detection, (Konishi, Yuille, Coughlan, & Zhu, 2003), or texture classification (Zhu & Mumford, 1997).

With selection of an appropriate generative model, ideal observer analysis allows inspection of a human observer's visual strategy and use of available visual information in a wide range of visual tasks. In a threshold pattern detection task that isolates early visual areas, ideal observer analysis has shown that the range of bandwidths of V1 receptive fields is matched to that of Gabor patches for which detection efficiency is highest (Kersten, 1984). Ideal observer analysis has been used, at intermediate levels, to study the information used by observers to recognize two-dimensional images of faces and letters (Gold, Murray, Bennett, & Sekuler, 2000; Pelli, Farell, & Moore, 2003), and, at higher levels, to constrain models of object recognition (Liu, Knill, & Kersten, 1995).

For the case of the giraffe, G, in front of the cluttered background, N, Bayes' rule provides the following equation:

$$P(G, N|I) \propto P(G, N)P(I - f(G, N))$$

The generative model provides the function $f(G, N)$ to generate the image, embodying rules of occlusion, illumination, perspective, etc. With no imaging noise, the likelihood is $\delta(I - f(G, N))$. The desired posterior—the probability of giraffe description G—would, in theory, be obtained by integrating out the shrub or background clutter variables; however, integrating out is easier said than done for cases other than low-dimensional additive noise cases, and developing models of natural clutter to estimate image signals is a current topic of interest (Zhu, Lanterman, & Miller, 1998). A maximum a posteriori ideal observer will pick giraffe description G for which the posterior probability is biggest. Manipulating either the shape of the modeled giraffe, or the characteristics of the background, can in principle test the prior assumptions that an observer uses in the analysis of the image. This approach will be most effective if the shape manipulation (the generative model incorporated in the ideal observer model) matches the built-in assumptions of the visual system (perhaps through the internal generative model, Kersten, Mamassian, & Yuille, 2004). In particular, the dimensions along which the generative model is varied should match the dimensions along which the visual system encodes object and image information. Further, the probabilities of generated images should match the observer's built-in expectations. These "expectations" may be hard-wired into early visual processing or may be learned. The actual process of selecting a model that matches the assumptions of the visual system is discussed below.

What are the strengths and limitations of the ideal observer approach to understanding human object perception? If thresholds for an object discrimination task are high, there can be two different causes: either the visual system is insensitive to the relevant information, or the information is intrinsically hard to extract because of the task. A strength of the ideal observer approach is that it quantifies the information available for a task and thereby allows separation of these two factors. For example, a recent study by Liu and Kersten showed that human visual sensitivity for discriminating three-dimensional asymmetric objects was better than for symmetric objects. But, because of the redundancy in the symmetric objects, accurate object information was intrinsically harder to extract: "harder" in the loose sense of less sensitivity because there were fewer independent image measurements for the symmetric objects. Therefore, when compared against ideal observer performance for the task, human discrimination performance was closer to ideal for symmetric objects (Liu & Kersten, 2003).

Clearly, the success of an ideal observer analysis hinges on the generative model used: more complete and accurate modeling of the real-world image formation process and real-world object priors (quantification of the available information) will lead to better estimates of the observer's priors. However, the computation of Bayesian estimates of object properties from real-world images and tasks is non-trivial, and this places limitations on ideal observer analysis. For example, categorization (e.g., giraffe vs. cat) is hard for Bayes, but easy for humans. The problem of "integrating out" multiple confounding variables from the posterior probability is computationally hard, and core to current progress in algorithms for Bayesian networks. It can be simpler to calculate an optimal estimate of the shape of an animal than to optimally categorize the animal, because the Bayes solution for categorization needs to solve the algorithmic problem of integrating over all the various poses and articulations that could be caused by a giraffe.

A second limitation is that, even apart from inference, specifying realistic generative models is a challenging problem. As discussed above, the problem has only been solved in a limited number of instances. One can measure local statistics from natural images, but using these to synthesize consistent global images is not straightforward for the same reason that a random synthesis of meaningful text is an unsolved problem. Measurements of local statistics provides the means to generate curve or text fragments that reflect naturally occurring probabilities; however, the random sampling of strings of such fragments do not necessarily produce likely or even consistent large scale structures such as an image of an "giraffe and bushes" or a new "F. Scott Fitzgerald novel" (Kersten et al., 2004). A realistic generative model therefore requires information about scene structure on several scales, accounting for the remarkable variability encountered in the real world.

1.2. Reverse correlation and classification images

Reverse correlation analysis, described in several other papers in this special issue, offers another approach to understanding the visual information used by an observer. Reverse correlation and classification image experiments seek an empirical estimate of the information used by an observer, whether neuronal receptive fields in reverse correlation studies (Ringach, Hawken, & Shapley, 2002), or templates used for visual tasks in classification image experiments (Gold et al., 2000). As with traditional applications of ideal observer analysis to vision, most reverse correlation studies assume a linear generative model, in this case one in which the explicit image information is modeled in two dimensions and masked by additive Gaussian noise:

$$f(S, N) = S_{2D} + N_{2D}$$

where S_{2D} is the image of the target object and N_{2D} is a two-dimensional matrix of pixel intensity variations with Gaussian distribution. This model has two primary consequences. First, it assumes that the relevant variables of interest (the explicit variables) are photometric, rather than geometric, since they are determined by their interaction with two-dimensional noise patterns. Second, the input parameter space for the generative model has the same dimensionality as the output, the images. However, as we show below, neither is an inherent or practical limitation.

It is worth discussing briefly the limitations of assuming a linear (photometric) image-based model for the object recognition process. A typical classification image experiment studying our test case of the giraffe in front of a background of bushes would use noise patterns, superimposed on the image, to probe which image features observers use to accomplish the discrimination task. The resulting classification image might reveal that noise patterns obscuring the neck or head interfere with the ability to judge the length of the neck and thereby distinguish between a giraffe and, say, a zebra. However, this result provides little information about the observer's ability to distinguish between a drinking giraffe and a zebra. The deformation from browsing giraffe to drinking giraffe is linear in a three-dimensional model, but could not be recovered from a two-dimensional classification image approach, which provides good information about illumination changes and moderate shape changes, but cannot be extrapolated to cover object motion or distortion on a larger scale. Another difficulty is non-linear interactions between object features: a long-necked animal will not look like a giraffe if the body of the animal is not also significantly tilted. This interaction between image features is problematic for all linear models, but particularly for image-based models, in which the dimensionality of the model space is so large that interactions between features such as neck length and body tilt are difficult to model. (For related criticisms, see Mangini and Biederman, this issue.)

The work in linear object classes suggests a simple extension of the classification image approach to object representation (e.g., Blanz & Vetter, 1999; Vetter & Poggio, 1997). The idea is to generate probe images by adding random deformations to the geometry of a prototype figure, and then to project the resulting figure onto two dimensions for display (e.g., similar to methods used by Liu & Kersten, 1998, but theirs was in the context of ideal observer analysis). Linearity is thus preserved in object space, but sacrificed in image intensity space (Jones & Poggio, 1995). Further, the dimensionality of the parameter (or signal) space is drastically reduced. However, care must be taken in defining the dimensions of the generative model, in order to minimize the impact of the bias unavoidably introduced by restricting the number of dimensions along which visual stimuli are varied.

1.3. From classification images to classification objects

The nature of neural representations is not fully understood, but in many cases it is possible to let neurophysiological studies guide the selection of a model that is well matched to studying visual responses. In the low level case, where there is a wealth of electrophysiological data, it has been shown that estimates of receptive field properties converge much more quickly when the noise patterns in a reverse correlation experiment are matched to response properties of neurons (Ringach, Sapiro, & Shapley, 1997). The selection of an appropriate generative model for studying visual responses in higher visual areas can be guided by what is known of visual tasks to which the animal is adapted and neural response properties in appropriate visual areas. While contrast is a strong modulator of activity in early visual areas, neural activity in inferior temporal areas is generally contrast invariant (Avidan et al., 2001). Rather than spatial location or orientation of line segments, neurons are tuned to various aspects of the shapes of object (Logothetis, 2000; Tanaka, 2003). And rather than a topology governed by location in visual space, the topology in the higher visual areas seems to be governed by spatial scale relative to the observer (Lerner, Hendler, Ben-Bashat, Harel, & Malach, 2001), or by more abstract categories of objects or shapes (Ishai et al., 1999; Tsunoda, Yamane, Nishizaki, & Tanifuji, 2001). Specialist regions of cortex are also found, which are most active during recognition of particularly salient categories of objects, such as faces (Kanwisher, McDermott, & Chun, 1997; Tarr & Gauthier, 2000). Thus, evidence points toward the existence of object models in higher visual areas that, in contrast to intensity-based receptive fields in lower visual areas, are based on geometric variations in the scene. Which geometric variations are important will depend on both the scene and the task under study.

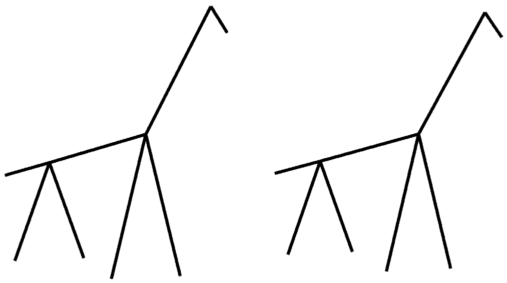

Fig. 1. As an example of the stimuli used, one observer's prototypical "giraffe" is shown as a stereo pair. (Left and right eyes' views are on the right and left, respectively.) Stimuli were not shown in stereo for the experiment.

2. Classification objects: an illustration

To illustrate the potential for using a three-dimensional generative model to study visual responses, we have adapted the classification image approach to show how one could characterize the human visual system's geometric representation of different classes of animals. Again, for the task of categorizing an animal, we will define a drastically simplified model of a generic quadruped, described by nine parameters,[3] modeled by line segments in three dimensions (see Fig. 1). The goal of the experiment is to understand how random deformations in the figure shape (modeled by a uniform distribution of the parameters defining the line segment vertices) affect the categorization of the animal. The conditions correspond to an assumed generative model,

$$I = f(A, N) = P_{2\mathrm{D}}R(A + N_{3\mathrm{D}})$$

where the rendering function is the product of $P_{2\mathrm{D}}$, an orthographic projection matrix, and R, a rotation matrix representing viewpoint. Vector A is unknown, representing the *xyz* coordinates connecting the key geometric features of the prototypical animal to be estimated, and N is the experimenter-generated deviation (random deformation) of the nine parameters from the prototypical values for a given species, in this case uniformly distributed within a cuboid. The similarity to the generative model described for classification image experiments shows that the difference between the two approaches is the addition of noise in the form of random deformations in the three-dimensional model, rather than noise patterns in the two-dimensional image. For the demonstration experiment, viewpoint is treated as a generic variable and thus also randomly varied. (Although we did not do this in our analysis below, the viewpoint coordinates could be treated as explicit variables, for example to produce an estimate of the canonical views of the animal.)

To demonstrate the kind of results and interpretation that can result from a geometry-based application of the reverse classification approach, we used sets of 300 presentations to estimate the internal representations for each of four species: cat, dog, horse, and giraffe. Each presentation involved a binary decision of whether the randomly generated and displayed animal looked like a named target animal. The classification object is calculated from the difference between the mean of the parameter sets (body tilt, neck length, body height, etc.) that resulted in a positive identification and the mean of all other trials (Fig. 2). Unlike classification

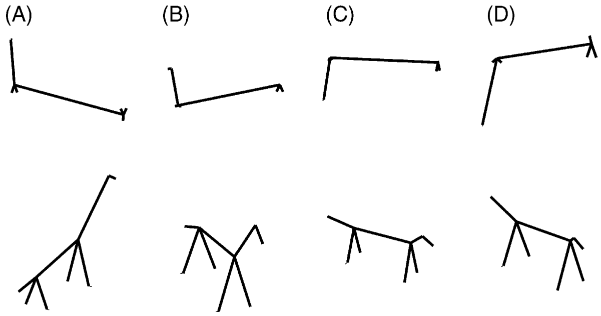

Fig. 2. 2D projections of the 3D classification objects estimated for four “species” of animals. (A) Giraffe; (B) horse; (C) dog, (D) cat. The first row shows the classification object generated from the mean of the model parameters resulting in a positive identification, minus the mean of the model parameters generating animals that did not look like the target animal. Each length is scaled relative to a unit body length. The second row shows the classification objects added to the mean animal (average of all trials tested for a given species), which produces a representation of the prototypical animal.

images, which provide intuitive two-dimensional maps highlighting image features that are key for object recognition or discrimination, the two-dimensional non-linearity of the classification object approach is obvious from the strange appearance of the classification objects. For example, the hind legs of the giraffe classification object are inverted. This is because body height is only slightly greater than the mean, but body tilt is much larger, resulting in inverted hind legs when displayed as an object. The correct interpretation of this result is that hind legs are slightly shorter than average; front legs are longer. The prototypical animal for each species is defined by adding the parameter values for the classification object to the mean parameter values from all trials. Prototypes were estimated for four species: giraffe, horse, dog, and cat.

The utility of each model attribute as a distinguishing feature can be judged by its ability to distinguish between species, and by the variability tolerated in the task. For example, body tilt is greater for the giraffe than for the other three animals; neck length is a feature that distinguishes all animals; tails at a more vertical angle are associated with cats (see Fig. 3). Neck angle and head angle had little to do with how the object is classified. The task-dependence of the distinction between explicit and generic variables is illustrated by this result. In the model, we treated the nine parameters as explicit variables, but several had little or no effect on the identification of the stick figures (body height, because of the scaling in image presentation, tail length, neck angle, head length, and head angle) and could have been treated as generic variables. The illustration of the drinking giraffe has already alluded to this point—the selection of model parameters, as well as the definition of explicit and generic variables, should rely on prior knowledge of both real world object arrangements and sensitivity of the visual system to information in the visual task being tested. In this particular study, viewpoint showed no interaction with the animal identification task: animals were generated with azimuth uniformly distributed from 0 to 360°, and elevation uniformly distributed from −30 to 30°, and both parameters showed the same uniform distribution for the subset of animals that were recognized as a particular species.

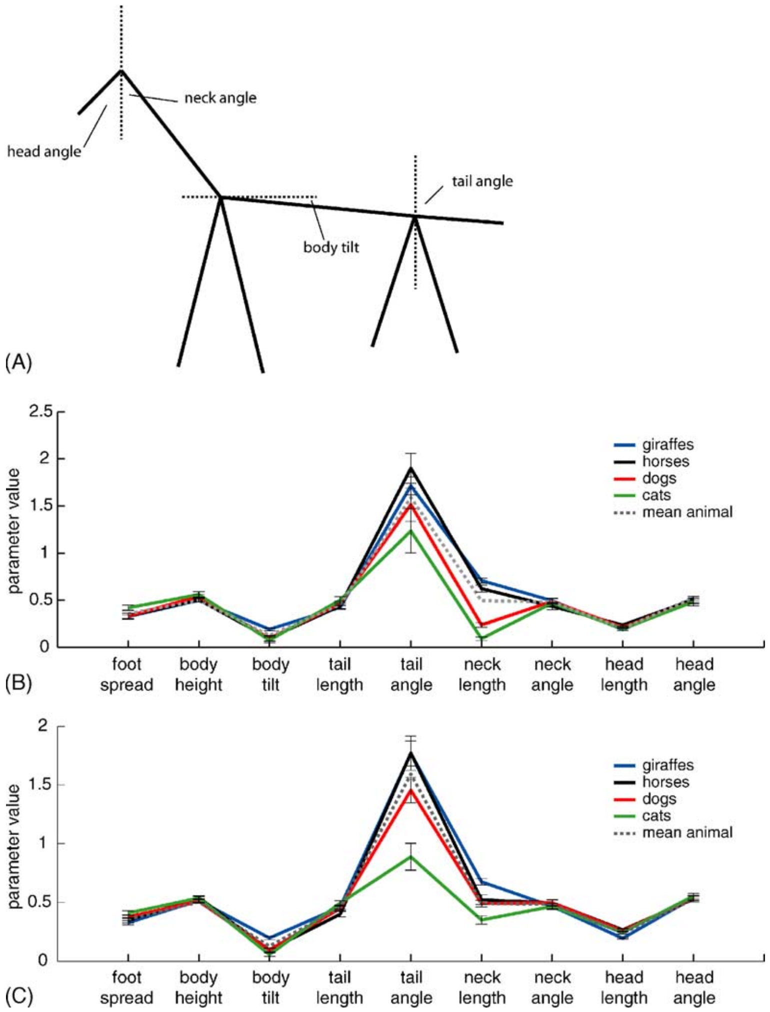

Fig. 3. Results of a classification object experiment. Stick figures were randomly generated from nine parameters; in 300 trials for each target animal, observers responded whether the stick figure looked like the named target animal. (A) Mean animal is shown with body part angles indicated. Mean parameter values are plotted for each prototype animal for one of the authors (B) and a naïve observer (C). Error bars indicate standard error of the mean. Dashed line indicates parameter values used to generate the mean animal. Parameter values are given either as lengths relative to body length or as angles in radians. When presented, each animal was scaled to fill a fixed window.

A natural extension of this process is to extend the decision from a binary decision (giraffe or not giraffe) to a decision in which the observer chooses between several different possible species (giraffe, horse, dog, cat, or not a recognizable animal). Developing the theory for such classification is beyond the scope of this paper. However, one can note that the results from such a task would depend on how the various classes are distributed in parameter space, and how the classification boundaries are maintained. For example, suppose classes A, B, C and D were distributed clock-wise on the corners of a square in a two-dimensional parameter space.

A line chosen to separate A from the set {B, C, D}, would be different from one chosen to separate A from B or A from C.

We ran an empirical test of a four-way classification task. Classification objects in this case were calculated as the difference between the mean parameters resulting in a positive identification as a particular species, minus the mean of all other trials, regardless of whether or not the stick figure was classified as a different animal or unrecognizable. The results (not shown) were essentially the same is in Fig. 3, the most significant difference being that the multiple-categorization task was more efficient, since data were being collected for all the prototypes simultaneously. The multiple-decision task may affect the probability that an observer will classify a particular stick figure as a recognizable animal. For example, an animal that looks like a cross between a horse and a giraffe may be rejected in either horse-only or giraffe-only trials, but will be classified as one or the other in multiple-decision trials. This would have the effect of broadening the range of parameters that is considered acceptable for generating a particular prototype animal.

3. Advantages and disadvantages of the classification object approach

The classification object approach probes the geometry of the internal representation by generating random deformations in the three-dimensional generative model, and measuring which aspects of the modeled noise most strongly affect classification of the object. For this demonstration, we used simple stick figures as an illustration, but the dimensionality of the parameter space could be increased (although not by much, compared to most applications of the classification image approach) to utilize more realistic three-dimensional geometric models (Cutzu & Edelman, 1998). Even a more complete model is a much smaller space than is typically used for classification image studies, yet is capable of preserving the full dimensionality of the image output space. A realistic generative model could include additional parameters to cover the full range of geometric variation (e.g., generalized cylinders, including non-rigid variations) as well as photometric components (e.g., shading).

For the specific case of face recognition, there are 3D databases that could be exploited with this technique. As in Leopold et al. (Leopold, O'Toole, Vetter, & Blanz, 2001), a face space can be created from an average face and linear interpolation between individuals. Then salient shape changes could be studied with a classification model approach, learning possible neural representations from the geometrical changes that most affect recognition or identification.

Although reverse correlation analysis is non-linear in its most general form (Marmarelis & Marmarelis, 1978), in that the generative model can approximate any polynomial function, it has the same limitations as any general purpose learning machine—the curse of dimensionality (Vapnik, 1995) and the bias-variance dilemma (Geman, Bienenstock, & Doursat, 1992). A generative model is used to reduce the dimensionality of the problem, and thereby the variance in the solution, but limiting the dimensionality of the model also increases the bias, or the probability that the solution systematically fails to fit the data. Minimizing bias requires as much data about the problem as is available. In our case, bias is introduced by limiting the range over which each parameter can vary, or by failing to parameterize a particular feature (such as body thickness or ear size) that is important in distinguishing between animals. Measurements of the actual range of features such as leg length, body tilt, or neck articulation in a set of images of real-world animals will aid in definition of a more appropriate model, reducing unwanted bias in the results.

The classification object approach described here also suffers from some of the same limitations discussed for the linear models used in classification image approaches. In the case where stick figures resembling both drinking giraffes and browsing giraffes elicit positive responses, the resulting prototype animal (linear combination of results) will have a very unusual articulation of the neck: horizontal, rather than drinking or browsing in the leaves. However, if this averaging were done in image space, the average would not be recognized as a giraffe at all. Likewise, the problem of interaction between features remains, but because the model is parametrically varied in a more appropriate space (with only a few dimensions), it is tractable to look at the "failed" stick figures, particularly those with long necks, and study how variations in other parameters (such as body tilt or height) affected identification of the animal as a giraffe.

In summary, the generative model provides the bridge between classification image approaches and ideal observer analysis. The classification object space corresponds to the signal together with a specification of the range of three-dimensional variations (e.g., animal prototype with random deformations), and the noise is defined by the generic variables (e.g., random viewpoints). Future studies should be able to use this common framework to test whether recognition efficiency, using ideal observer analysis, can be predicted from the probabilistic object models (means of prototypes plus ranges of deformations, e.g., Fig. 3) discovered using classification objects.

Notes

1. If there is no imaging noise, the distribution, $P(I - f(S, N))$ is a delta function, $\delta(I - f(S, N))$. The delta distribution is a zero-tolerance filter for wrong scenes—it is zero for scenes that don't predict the image data and uniformly high for those that do.
2. More generally, Bayesian decision theory softens the sharp distinction between explicit and generic (noise) variables by defining a loss function $L(\Sigma, S)$ which is the penalty for Σ (the estimate of S) when the true parameter is S. Then the optimal decision minimizes the risk:
$$R(\Sigma_N, \Sigma_S) = \sum_{N,S} L(N, \Sigma_N; S, \Sigma_S) P(S, N|I)$$
With a loss function, $-\delta(\Sigma_S - S)$, where the cost to all errors in the generic variable equal, minimizing risk is equivalent to summing the posterior with respect to the generic variable,
$$P(S|I) = \sum_{N} P(N, S|I)$$
and choosing the maximum of this posterior distribution (Geisler & Kersten, 2002; Liu, Kersten, & Knill, 1999).
3. All four feet are constrained to be on the ground; the animal is symmetric across the plane perpendicular to the ground and running through the body axis. We chose fixed

ranges for each of the model parameters: body tilt, body height (distance from center of body segment to the plane of the feet), foot spread, tail length, tail angle, neck length, neck angle, head length and head angle.

Acknowledgments

This work was supported by NIH RO1 EY02587 and NSF/IGERT DGE 9870633.

References

Avidan, G., Harel, M., Hendler, T., Ben-Bashat, D., Zohary, E., & Malach, R. (2001). Contrast sensitivity in human visual areas and its relationship to object recognition. *J. Neurophysiol.*, *87*, 3112–3116.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *SIGGRAPH'99 conference proceedings* (pp. 187–194). Los Angeles, CA, USA.

Cutzu, F., & Edelman, S. (1998). Representation of object similarity in human vision: Psychophysics and a computational model. *Vision Research*, *38*, 2229–2257.

Geisler, W. S., & Albrecht, D. G. (1995). Bayesian analysis of identification performance in monkey visual cortex: Nonlinear mechanisms and stimulus certainty. *Vision Research*, *35*(19), 2723–2730.

Geisler, W. S., & Kersten, D. (2002). Illusions, perception and bayes. *Nature Neuroscience*, *5*(6), 508–510.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology*, *10*, 1663–1666.

Grenander, U. (1996). *Elements of pattern theory*. Baltimore: Johns Hopkins University Press.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences, USA*, *96*, 9370–9384.

Jones, M. J., & Poggio, T. (1995). Model-based matching of line drawings by linear combinations of prototypes. *Proceedings of the IEEE 5th International Conference on computer vision* (pp. 531–536). Cambridge, MA: IEEE Computer Society Press.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, *17*(11), 4302–4311.

Kersten, D. (1984). Spatial summation in visual noise. *Vision Research*, *24*(12), 1977–1990.

Kersten, D. (1987). Predictability and redundancy of natural images. *Journal of the Optical Society of America A*, *4*, 2395–2400.

Kersten, D. (2002). Object perception: Generative image models and Bayesian inference. In H. H. Bülthoff, S.-W. Lee, T. Poggio, & C. Wallraven (Eds.), *Biologically motivated computer vision 2002* (pp. 207–218). Tübingen, Germany: Springer-Verlag, Berlin Heidelberg.

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

Konishi, S. M., Yuille, A. L., Coughlan, J. M., & Zhu, S. C. (2003). Statistical edge detection: Learning and evaluating edge cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *25*(1), 57–74.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*(1), 89–94.

Lerner, Y., Hendler, T., Ben-Bashat, D., Harel, M., & Malach, R. (2001). A hierarchical axis of object processing stages in the human visual cortex. *Cerebral Cortex*, *11*, 287–297.

Liu, Z., & Kersten, D. (1998). 2D observers for human 3D object recognition? *Vision Research*, *38*(15/16), 2507–2519.

Liu, Z., & Kersten, D. (2003). Three-dimensional symmetric shapes are discriminated more efficiently than asymmetric ones. *Journal of the Optical Society of America A*, *20*(7), 1331–1340.

Liu, Z., Kersten, D., & Knill, D. C. (1999). Dissociating stimulus information from internal representation—a case study in object recognition. *Vision Research*, *39*(3), 603–612.

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

Logothetis, N. K. (2000). Object recognition: Holistic representations in the monkey brain. *Spatial Vision*, *13*(2/3), 165–178.

Marmarelis, P. Z., & Marmarelis, V. Z. (1978). *Analysis of physiological systems: The white-noise approach*. New York: Plenum Press.

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature*, *423*, 752–756.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature*, *317*, 314–319.

Richards, V. M., & Zhu, S. (1994). Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients. *Journal of the Acoustical Society of America*, *95*(1), 423–434.

Ringach, D. L., Hawken, M. J., & Shapley, R. M. (2002). Receptive field structure of neurons in monkey primary visual cortex revealed by stimulation with natural image sequences. *Journal of Vision*, *2*(1), 12–24.

Ringach, D., Sapiro, G., & Shapley, R. M. (1997). A subspace reverse-correlation technique for the study of visual neurons. *Vision Research*, *37*(17), 2455–2464.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities. *Cerebral Cortex*, *13*(1), 90–99.

Tarr, M. J., & Gauthier, I. (2000). FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, *3*(8), 764–769.

Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, *2*, 371–387.

Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, *4*(8), 832–838.

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

Vapnik, V. N. (1995). *The nature of statistical learning*. New York: Springer-Verlag.

Vetter, T., & Poggio, T. (1997). Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 733–742.

Watson, A. B., Barlow, H. B., & Robson, J. G. (1983). What does the eye see best? *Nature*, *31*, 419–422.

Zhu, S., Lanterman, A. D., & Miller, M. I. (1998). Clutter modeling and performance analysis in automatic target recognition. *Workshop on detection and classification of difficult targets*. Redstone Arsenal, AL.

Zhu, S. C., & Mumford, D. (1997). Learning generic prior models for visual computation. *Proceedings of the Conference on computer vision and pattern recognition* (p. 463). Puerto Rico.