2D Observers in 3D Object Recognition

Zili Liu
NEC Research Institute
Princeton, NJ 08540

Daniel Kersten
University of Minnesota
Minneapolis, MN 55455

Abstract

Converging evidence has shown that human object recognition depends on the observers' familiarity with objects' appearance. The more similar the objects are, the stronger this dependence is, and the more important two-dimensional (2D) image information is to discriminate these objects from one another. The degree to which 3D structural information is used, however, still remains an area of strong debate. Previously, we showed that all models that allow rotations in the image plane of independent 2D templates could not account for human performance in discriminating novel object views as a result of 3D rotations. We now present results from models of generalized radial basis functions (GRBF), 2D closest template matching that allows 2D affine transformations of independent 2D templates, and Bayesian statistical estimator that integrates over all possible 2D affine transformations. The performance of the human observers *relative* to each of the models is better for the novel views than for the learned template views, a result unpredicted if the human observers employed up to 2D affine transformations to independent 2D templates. The Bayesian estimator yields provably the optimal performance among all models of 2D affine transformations with independent 2D templates. Therefore, no models of 2D affine operations with independent 2D templates account for the human observers' performance. We suggest that the human observers used 3D structural information of the objects, which is also supported by the improved performance as the objects' 3D structural regularity increased.

1 Introduction

Object recognition is one of the most important functions in human vision. To understand human object recognition, it is essential to understand the nature of human object representations in memory. By definition, object recognition is the matching of an object's representation with an input object image. Therefore, in any object recognition studies, the nature of the object representation has to be inferred from the recognition performance, by taking into account the contribution from the image information. An ideal observer is precisely such a computational model in need whose recognition performance is restricted only by the available image information and is otherwise optimal, in signal detection sense, irrespective of how the model is implemented.

To illustrate the importance of image information normalization, let us look at the recent debate in human object recognition on the dependence of the representation on viewpoint variation (Biederman & Gerhardstein, 1995; Tarr & Bülthoff, 1995) . The central criterion, assumed by all in the debate, is the equivalence in viewpoint dependence between the representation in memory and the recognition performance. In other words, the assumption is that a viewpoint dependent performance implies a viewpoint dependent representation, and that a viewpoint independent performance a viewpoint independent representation. However, given that any recognition performance depends on the input image information, which is necessarily viewpoint dependent, the viewpoint dependence of recognition performance is logically neither necessary nor sufficient for the viewpoint dependence of the representation. Image information has to be factored out first.

The second aspect of an ideal observer is that it is implementation free. To illustrate its importance in hypothesis testing in object recognition, let us look at the Generalized Radial Basis Functions model (GRBF) (Poggio & Edelman, 1990), as compared with human object recognition. The model stores a number of 2D templates $\{\mathbf{T}_i\}$ of a 3D object $\mathbf{O}$, and recognizes or rejects a stimulus image $\mathbf{S}$ by the following similarity measure $\Sigma_i c_i \exp\left(-\frac{\|\mathbf{T}_i-\mathbf{S}\|^2}{2\sigma^2}\right)$, where c_i and σ are constants. The mimicry of the model's performance as a function of viewpoint to that of human observers has led to the claim that the human visual system may indeed, as does the model, use 2D stored views to recognize 3D objects (Bülthoff & Edelman, 1992). Such a conclusion, however, overlooks implementational constraints in the model, because the model's performance also depends on its implementations. Conceivably, a model with some 3D information of the objects can also mimic human performance, so long as it is appropriately implemented.

In contrast, an ideal observer computes the optimal performance that is only limited by the stimulus information and explicitly specified assumptions. It therefore yields the best possible performance among the class of models with the same stimulus input and assumptions. We are particularly interested in constrained ideal observers that are restricted in functionally significant aspects (e.g., a constrained 2D ideal observer that stores independent 2D templates and has access only to 2D affine transformations). The key idea is that such constrained ideal observer is the best in its class, so if humans outperform this ideal, they must have used more than what is available to the ideal. The constrained ideal and its whole class therefore cannot account for human performance. Thus, this lower-bound approach provides

a powerful and rigorous tool of hypothesis testing for human object recognition.

2 The observers

In this section, we will first define an experimental task, in which the computational models yield provably the best possible performance under their specified conditions. We will then review the 2D ideal and GRBF observers derived in (Liu, Knill, & Kersten, 1995) , the 2D affine matching model in (Werman & Weinshall, 1995), and derive a Bayesian 2D affine ideal observer. In the next section, we will compare human observers' performance in (Liu, Knill, & Kersten, 1995) with these models. We predict that if humans outperform these 2D computational models, then humans must have used more than what is available to the models.

Let us first define the task. The observer (human or computational) first look at the 2D images of a wire frame object from a number of viewpoints. These images will be called templates $\{\mathbf{T}_i\}$. Then two distorted versions of the original 3D object are displayed. They are obtained by adding 3D Gaussian positional noise (i.i.d.) to the vertices of the original object. One distorted object is called the target, whose Gaussian noise has a constant variance. The other is the distractor, whose noise has a larger, variable variance. The two objects are displayed from the same viewpoint, which can be either from one of the template views, or a novel view due to 3D rotation. The task is to choose the one that is more similar to the learned original object. The observer's performance is measured by the threshold variance that gives rise to 75% correct performance. The optimal strategy is to choose the stimulus $\mathbf{S}$ with a larger probability $p(\mathbf{O}|\mathbf{S})$. From the Bayes rule, this is to choose the larger $p(\mathbf{S}|\mathbf{O})$.

Assume that no model can reconstruct the 3D structure of the object from its independent templates $\{\mathbf{T}_i\}$. Assume also that the prior probability $p(\mathbf{T}_i)$ is constant. Let $\mathbf{S}$ and $\mathbf{T}_i$ be represented by their (x, y) vertex coordinates: $\begin{pmatrix} \mathbf{X} & \mathbf{Y} \end{pmatrix}^T$, where $\mathbf{X} = \left(x^1, x^2, \ldots, x^n\right)$, $\mathbf{Y} = \left(y^1, y^2, \ldots, y^n\right)$. We assume that the correspondence between $\mathbf{S}$ and $\mathbf{T}_i$ is solved up to a reflection ambiguity, which is equivalent to an additional template: $\mathbf{T}_i^r = \begin{pmatrix} \mathbf{X}^r & \mathbf{Y}^r \end{pmatrix}^T$, where $\mathbf{X}^r = \left(x^n, \ldots, x^2, x^1\right)$, $\mathbf{Y}^r = \left(y^n, \ldots, y^2, y^1\right)$. We still denote the template set as $\{\mathbf{T}_i\}$. Therefore,

$$p(\mathbf{S}|\mathbf{O}) = \Sigma p(\mathbf{S}|\mathbf{T}_i)p(\mathbf{T}_i). \tag{1}$$

In what follows, we will compute $p(\mathbf{S}|\mathbf{T}_i)p(\mathbf{T}_i)$, with the assumption that $\mathbf{S} = \mathcal{F}(\mathbf{T}_i) + \mathbf{N}(\mathbf{0}, \sigma\mathbf{I}_{2n})$, where $\mathbf{N}$ is the Gaussian distribution, $\mathbf{I}_{2n}$ the $2n \times 2n$ identity matrix, and $\mathcal{F}$ a 2D transformation. For the 2D ideal observer, $\mathcal{F}$ is a rigid 2D rotation. For the GRBF model, $\mathcal{F}$ assigns a linear coefficient to each template $\mathbf{T}_i$, in addition to a 2D rotation. For the 2D affine matching model, $\mathcal{F}$ represents the 2D affine transformation that minimizes $\|\mathbf{S} - \mathbf{T}_i\|^2$, after $\mathbf{S}$ and $\mathbf{T}_i$ are normalized in size. For the 2D affine ideal observer, $\mathcal{F}$ represents all possible 2D affine transformations applicable to $\mathbf{T}_i$.

2.1 The 2D ideal

The templates are the original 2D images, their mirror reflections, and 2D rotations (in angle ϕ) in the image plane. Assume that the stimulus $\mathbf{S}$ is generated by adding

Gaussian noise to the templates. The probability $p(\mathbf{S}|\mathbf{O})$ is an integration over all templates and their reflections and rotations. The detailed derivation for the 2D ideal and the GRBF model can be found in (Liu, Knill, & Kersten, 1995).

$$\Sigma p(\mathbf{S}|\mathbf{T}_i)p(\mathbf{T}_i) \propto \Sigma \int d\phi \exp\left(-\frac{\|\mathbf{S}-\mathbf{T}_i(\phi)\|^2}{2\sigma^2}\right). \tag{2}$$

2.2 The GRBF model

The model has the same template set as the 2D ideal does. Its training requires that

$$\Sigma_i \int_0^{2\pi} d\phi c_i(\phi) N(\|\mathbf{T}_j - \mathbf{T}_i(\phi)\|, \sigma) = 1, j = 1, 2, \ldots, \tag{3}$$

with which $\{c_i\}$ can be obtained optimally in the least square sense. When a pair of new stimuli $\{\mathbf{S}\}$ come in, the optimal decision is to choose the one closer to the learned prototype, in other words, the one with a smaller value of

$$\|1 - \Sigma \int_0^{2\pi} d\phi c_i(\phi) \exp\left(-\frac{\|\mathbf{S}-\mathbf{T}_i(\phi)\|^2}{2\sigma^2}\right)\|. \tag{4}$$

2.3 The 2D affine matching model

It has been proved (Liu, Knill, & Kersten, 1995) that the smallest Euclidean distance $D(\mathbf{S}, \mathbf{T})$ between $\mathbf{S}$ and $\mathbf{T}$ is, when $\mathbf{T}$ is allowed up to a 2D affine transformation, and after a scale normalization $\mathbf{S} \rightarrow \frac{\mathbf{S}}{\|\mathbf{S}\|}$, $\mathbf{T} \rightarrow \frac{\mathbf{T}}{\|\mathbf{T}\|}$,

$$D^2(\mathbf{S}, \mathbf{T}) = 1 - \frac{tr(\mathbf{S}^+\mathbf{S} \cdot \mathbf{T}^T\mathbf{T})}{\|\mathbf{T}\|^2}, \tag{5}$$

where tr strands for *trace*, and $\mathbf{S}^+ = \mathbf{S}^T(\mathbf{S}\mathbf{S}^T)^{-1}$. The optimal strategy, therefore, is to choose the $\mathbf{S}$ that gives rise to the larger of $\Sigma \exp\left(-\frac{D^2(\mathbf{S},\mathbf{T}_i)}{2\sigma^2}\right)$, or the smaller of $\Sigma D^2(\mathbf{S}, \mathbf{T}_i)$. (Since no probability is defined in this model, both measures will be used and the results from the better one will be reported.)

2.4 The 2D affine ideal

We now calculate the Bayesian probability by assuming that the prior probability distribution of the affine transformation, which is applied to the template $\mathbf{T}_i$, $\mathbf{AT} + \mathbf{T_r} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mathbf{T}_i + \begin{pmatrix} t_x & \ldots & t_x \\ t_y & \ldots & t_y \end{pmatrix}$, obeys a Gaussian distribution $\mathbf{N}(\mathbf{X_0}, \mu\mathbf{I}_6)$, where $\mathbf{X}_0$ is the identity transformation $\mathbf{X}_0 = (a, b, c, d, t_x, t_y)^T = (1, 0, 0, 1, 0, 0)^T$. We have

$$\begin{aligned} \Sigma p(\mathbf{S}|\mathbf{T}_i) &= \Sigma \int_{-\infty}^{\infty} dadbdcdd\, dt_x dt_y \exp\left(-\frac{\|\mathbf{AT}_i + \mathbf{T_r} - \mathbf{S}\|^2}{2\sigma^2}\right) & (6) \\ &= \Sigma \frac{C(n, \sigma, \mu)}{det\,(\mathbf{Q}_i')} \exp\left(\frac{tr\left(\mathbf{K}_i^T\mathbf{Q}_i(\mathbf{Q}_i')^{-1}\mathbf{Q}_i\mathbf{K}_i\right)}{2\sigma^2}\right), & (7) \end{aligned}$$

where $C(n, \sigma, \mu)$ is a function of n, σ, μ; $\mathbf{Q}' = \mathbf{Q} + \mu^{-2}\mathbf{I}_2$, and

$$\mathbf{Q} = \begin{pmatrix} \mathbf{X_T} \cdot \mathbf{X_T} & \mathbf{X_T} \cdot \mathbf{Y_T} \\ \mathbf{Y_T} \cdot \mathbf{X_T} & \mathbf{Y_T} \cdot \mathbf{Y_T} \end{pmatrix}, \mathbf{QK} = \begin{pmatrix} \mathbf{X_T} \cdot \mathbf{X_S} & \mathbf{Y_T} \cdot \mathbf{X_S} \\ \mathbf{X_T} \cdot \mathbf{Y_S} & \mathbf{Y_T} \cdot \mathbf{Y_S} \end{pmatrix} + \mu^{-2}\mathbf{I}_2. \quad (8)$$

The free parameters are μ and the number of 2D rotated copies for each $\mathbf{T}_i$ (since a 2D affine transformation implicitly includes 2D rotations, and since a specific prior probability distribution $\mathbf{N}(\mathbf{X}_0, \mu\mathbf{I})$ is assumed, both free parameters should be explored together to search for the optimal results).

Figure 1: Stimulus classes with the increasing structural regularity: Balls, Irregular, Symmetric, and V-Shaped. Three objects in each class were tested in the experiment.

2.5 The human observers

Three naive subjects were tested with four classes of objects: Balls, Irregular, Symmetric, and V-Shaped (Fig. 1), three objects each. For each object, 11 template views were learned by rotating the object 60°/step, around the X- and Y-axis, respectively. The 2D images were generated by parallel projection, and viewed monocularly. The viewing distance was 1.5 m. During the test, the standard deviation of the Gaussian noise added to the target object was $\sigma_t = 0.254$ cm. No feedback was provided.

Because the image information available to the humans was more than what was available to the models (shading and occlusion in addition to the (x, y) positions of the vertices), both learned and novel views were tested in a randomly interleaved fashion. Therefore, the strategy that humans used in the task for the learned and novel views should be the same. The number of self-occlusions, which in principle provided relative depth information, was counted and was about equal in both learned and novel view conditions. The shading information was also likely to be equal for the learned and novel views. Therefore, these additional information should be about equal for the learned and novel views, and should not affect the comparison between the performance (humans relative to a model) for the learned vs. for the novel views. We predict that if the humans used, say, up to a 2D affine strategy, then their performance *relative* to the 2D affine ideal should not be higher for the novel views than for the learned views. Otherwise, humans must have used more than what is available to the 2D affine ideal. One possibility would be that the humans do not use the stored templates independently, but reconstruct partial 3D structure from the templates, and use 3D knowledge of the object in discriminating novel views. One reason to use the four classes of objects with increasing structural regularity is that the structural regularity is a 3D property (e.g., 3D Symmetric vs. Irregular), which the 2D models cannot capture (the only exception is the planar V-Shaped objects, for which the 2D affine models completely capture 3D rotations, and are therefore the "correct" models. The V-Shaped objects

were used in the 2D affine case as a benchmark). If humans' performance improves with the increasing structural regularity of the objects, this would lend additional support for the hypothesis that humans have used 3D information in the task.

3 Results

A stair-case procedure (Waston & Pelli, 1983) was used to track the observers' performance at 75% correct level for the learned and novel views, respectively, 120 trials for the humans, and 2000 trials for each of the models. For the GRBF model, the standard deviation of the Gaussian function was also sampled to search for the best result for the novel views for *each* of the 12 objects, the result for the learned views was obtained under the same standard deviation for that particular object. This resulted in a conservative hypothesis testing for the following reasons: (1) Since no feedback was provided in the human experiment and the learned and novel views were randomly intermixed, it is not straightforward for the model to find the best standard deviation for the novel views, particularly because the best standard deviation for the novel views was not the same as that for the learned. The performance for the novel views is therefore the upper limit of the model's performance. (2) The subjects' performance relative to the model will be defined as statistical efficiency. The above method will yield the lowest possible efficiency for the novel views, and a higher efficiency for the learned views, since the best standard deviation for the novel views is different from that for the learned views. Because our hypothesis depends on a higher statistical efficiency for the novel views than for the learned views, this method will make such a putative difference even smaller. Likewise, for the 2D affine ideal, the number of 2D rotated copies of each template $\mathbf{T}_i$ and the value μ were both extensively sampled, and the best performance for the novel views was selected accordingly. The result for the learned views corresponding to the same parameters was selected. This choice also makes it a conservative hypothesis testing. Fig. 2 shows the threshold performance, the standard deviation of the Gaussian noise added to the distractor to keep the 75% correct performance, of the human observers and the models. Note that both the 2D affine matching model and the 2D affine ideal observer yielded the same performance for the learned and novel views, for the planar V-Shaped objects.

In order to compare human relative to an ideal observer's performance, statistical efficiency $\mathcal{E}$ is defined (Fisher, 1925) that quantifies the relative information used by humans relative to the ideal observer: $\mathcal{E} = \left(\frac{d'_{human}}{d'_{ideal}}\right)^2$, where d' is the discrimination index. We have shown in (Liu, Knill, & Kersten, 1995) that, in our task,

$$\mathcal{E} = \frac{\left(\sigma_d^{ideal}\right)^2 - (\sigma_t)^2}{\left(\sigma_d^{human}\right)^2 - (\sigma_t)^2}, \tag{9}$$

where σ is the threshold — the standard deviation of the Gaussian noise, σ_t is that added to the target, and σ_d to the distractor. Fig. 3 shows the statistical efficiency of the human observers relative to each of the four models.

We note in Fig. 3 that the efficiency for the novel views are higher than that for the learned views (several of them even exceeded 100%), except for the planar V-Shaped objects. We are particularly interested in the Irregular and Symmetric objects in the 2D affine ideal case, in which the pairwise comparison between the learned

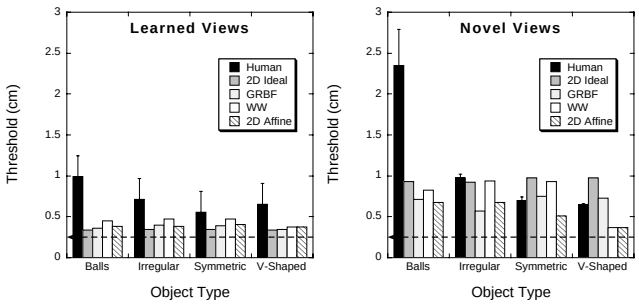

Figure 2: The threshold standard deviation of the Gaussian noise added to the distractor in the test pair that keeps an observer's performance at 75% correct level, for the learned and novel views, respectively. The dotted line is the standard deviation of the Gaussian noise added to the target in the test pair. The error bars for the human observers represent standard errors.

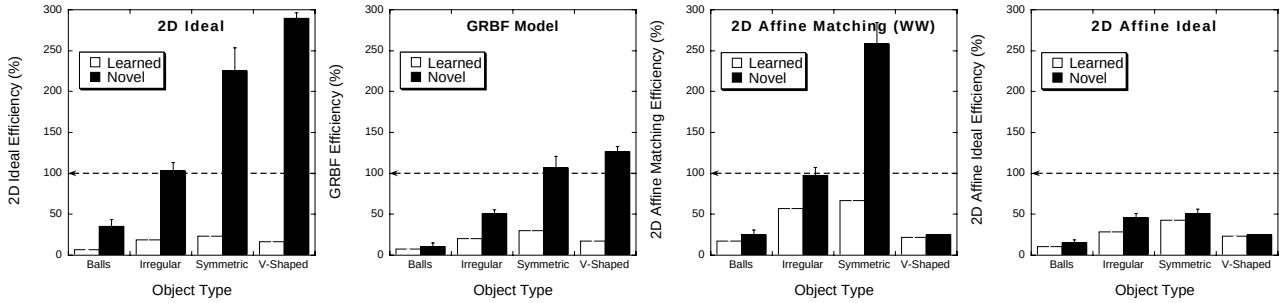

Figure 3: Statistical efficiencies of human observers relative to the 2D ideal observer, the GRBF model, the 2D affine best matching model of Werman & Weinshall (1995), and the 2D affine ideal observer. The error bars are standard errors.

and novel views across the six objects and three observers yielded a significant difference (binomial, $p < .05$). This suggests that the 2D affine ideal observer cannot account for the human observers' performance, because if the human observers used up to a 2D linear template matching strategy, their relative performance for the novel views cannot be better than for the learned views. We speculate therefore that 3D information was used by the human observers (e.g., 3D symmetry). This is supported in addition by the increasing efficiencies as the structural regularity increased from the Balls, Irregular, to Symmetric objects (except for the V-Shaped objects with 2D affine models).

The reason that, for example, the 2D affine ideal observer still cannot account for human observers' performance even with its additional power and flexibility relative to the 2D ideal observer, which only allows for 2D rotations, can be explained as follows. A 2D affine transformation is still an approximation, rather than a precise model, of a 3D object's 3D rotation. Therefore, although the ideal has a larger range of transformations applicable to a template, it at the same time also creates many more useless templates that are more similar to the distractor than to the target, which only impedes the model's performance. Therefore, by having simultaneously a more powerful and flexible model and more false matches, it is not obvious why

the eventual performance can explain away completely the human data. The results from the two 2D affine models — the closest match and the Bayesian models, show that they still cannot.

4 Conclusions

Computational models of visual cognition are often subject to information theoretic as well as implementational constraints. When a model's performance mimics that of human observers, it is difficult to interpret which aspects of the model, if any, characterizes the human visual system. For example, human object recognition could be simulated by both a GRBF model and a model with partial 3D information of the object. The approach we are advocating here is that, instead of trying to mimic human performance by a computational model, we design an implementation free model that yields the best possible performance under explicitly specified computational constraints. This model serves as a rigorous benchmark, and if human observers outperform it, we can conclude firmly that the humans must have used better computational strategies than the model can. For instance, models of independent 2D templates with 2D linear operations cannot account for the human performance, suggesting that the humans may have used the templates to reconstruct (crude) 3D structure of the object. This result is difficult to conceive without the ideal observer analysis.

References

[1] Biederman I and Gerhardstein P C. Viewpoint dependent mechanisms in visual object recognition: a critical analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 21:1506–1514, 1995.

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

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

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

[5] Liu Z, Knill D C, and Kersten D. Object classification for human and ideal observers. *Vision Research*, 35:549–568, 1995.

[6] Werman M and Weinshall D. Similarity and affine invariant distances between 2D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:810–814, 1995.

[7] Watson A B and Pelli D G. QUEST: A Bayesian adaptive psychometric method. *Perception and Psychophysics*, 33:113–120, 1983.

[8] Fisher R A. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.