PERGAMON

Vision Research 38 (1998) 2507–2519

Vision Research

2D observers for human 3D object recognition?

Zili Liu [a,*], Daniel Kersten [b]

[a] NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, USA
[b] Department of Psychology, University of Minnesota, Minneapolis, MN 55455, USA

Received 5 February 1997; received in revised form 20 January 1998

Abstract

In human object recognition, converging evidence has shown that subjects' performance depends on their familiarity with an object's appearance. The extent of such dependence is a function of the inter-object similarity. The more similar the objects are, the stronger this dependence will be and the more dominant the two-dimensional (2D) image-based information will be. However, the degree to which three-dimensional (3D) model-based information is used remains an area of strong debate. Previously the authors showed that all models with independent 2D templates that allowed 2D rotations in the image plane cannot account for human performance in discriminating novel object views [1]. Here the authors derive an analytic formulation of a Bayesian model that gives rise to the best possible performance under 2D affine transformations and demonstrate that this model cannot account for human performance in 3D object discrimination. Relative to this model, human statistical efficiency is higher for novel views than for learned views, suggesting that human observers have used some 3D structural information.

Keywords: Affine transformation; Object recognition; Object representation; Ideal observer; Template matching

1. Introduction

A basic component in three-dimensional (3D) object recognition is a process that matches the input stimulus to stored object representations in memory. The search for a match, when viewpoint invariant features (e.g. color or material) are absent, must be based on the object shape. A major challenge for object recognition is to understand how potential matches are verified despite shape variations in the image due to rotations in viewpoint.

Empirical evidence has shown that human object recognition strongly depends on familiar views, a result particularly pronounced for structurally similar objects [2–5]. These studies leave open, however, the question of how much 3D information contributes to object recognition. In contrast, empirical evidence in support of 3D model-based recognition suggests that object recognition is viewpoint dependent only when major object components disappear and new components come into view, for structurally dissimilar objects [6][1]. The studies do not, however, resolve the possibility that since the objects are dissimilar, a two-dimensional (2D) based qualitative representation already suffices to distinguish an object from the rest within a large range of viewpoint change.

1.1. View-approximation models

To clarify what the authors mean by 2D versus 3D information, let us consider one class of models for shape-based recognition, which the authors refer to as view-approximation models. (The authors postpone consideration of the more powerful view-combination models to the Discussion. See [7] for a general discussion of various classes of object recognition models.) View-approximation models assume that views are arbitrary samples, whose only link is a common label (e.g. the name of the object). These views have come to be associated with each other through experience. Thus, such models are inherently viewpoint dependent. For example, assume that an object is represented by two independent views. The task is to decide whether a novel view belongs to the object. The strong version of view-approximation maintains that in order to recog-

* Corresponding author. Fax: +1 609 9512481; e-mail: zliu@research.nj.nec.com.

[1] But see [24]; [25].

0042-6989/98/$19.00
PII: S0042-6989(98)00063-7

nize a novel view, a similarity measure is calculated independently between this view and each of the two stored views [8,4]. Recognition is a function of these measurements. The simplest function is the nearest neighbor scheme, where a match is based on the closest view in memory. A more sophisticated scheme is the Bayes classifier that combines the evidence over the collection of views optimally.

A more flexible version of view-approximation is to allow, in addition to combinations of the similarities, transformations on each stored view. For example, a novel 2D view can be translated and rotated in the 2D image plane before matching with each of the stored 2D views. [1] showed that human observers exceeded even the optimal model that used this strategy (which the authors referred to as a '2D/2D ideal observer' (2D model/2D input)). Thus, the results excluded both the strong and more flexible models above. The authors did not, however, exclude view-approximation models with even more flexible transformations. One example of the 2D/2D observer class is to allow 2D affine transformations to each of the templates before similarity computations. A 2D affine transformation is any linear transformation that includes translation, rotation, scaling and stretching in the image plane, which the authors will define shortly. Such a 2D affine transformation exactly characterizes 3D rotations of 2D planar objects under orthographic projection (see [9] for a summary[2]) and approximates, in a small range, depth rotations of 3D objects [10]. The primary purpose of this paper is to test whether 2D affine transformations account for human performance for 3D object recognition.

1.2. Distinguishing models experimentally: the ideal observer approach

The authors approach is to first construct a 2D affine model that gives rise to the best possible performance, which the authors call the 2D affine ideal observer. The authors then test whether this ideal observer accounts for human performance or not. If not, the authors can reject this ideal observer and all the models suboptimal to it, as models for human object recognition.

In the following, the authors first derive the 2D affine ideal observer. For quantitative comparison, the authors describe three additional models that have been proposed in the literature. First, the authors introduce the model by [11] that matches two point sets using 2D affine transformations. This model provides a particularly simple approximation to the 2D affine ideal observer. Second, the authors introduce a model by [12] that recognizes a 2D image of a set of 3D points from a single 2D template. Third, in order to compare with the Generalized Radial Basis Functions (GRBF) model in [1], the authors present an improved GRBF model that adjusts the variance of its radial basis (Gaussian) functions to search for the best result. Finally, the authors compare human performance with these models in a 3D object discrimination task [1]. The task requires observers to discriminate which of two objects is more similar to a learned object. The task provides a straightforward way of measuring the efficiency of the human matching process for novel object views. The authors use wire objects as the stimuli because they are the simplest objects that obey the assumptions of these models.

[2] When a planar object is rotated in depth under orthographic projection, the object is scaled in the image plane along the direction perpendicular to the rotational axis. A 2D affine transformation can also scale a 2D image along one direction. That is why 2D affine transformation can exactly characterize any 3D rotation of a 2D planar object.

2. The computational models

In order to provide a clear context of what the computational models are supposed to do, the authors briefly describe the task that both the human observers and models face [1]. The objects are bent wires whose vertex feature points are assumed visible from all viewing angles with known correspondence (i.e. the feature points are labeled). An image of an object is represented by the (x, y) coordinates of its feature points. The object (termed prototype) is first learned from a number of its images. Then a pair of objects are generated from this prototype by adding independent 3D positional Gaussian noise at the feature 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. The task is to choose from the two an object that is more similar, in Euclidean distance of the feature points, to the prototype object.

2.1. The 2D affine ideal observer

Here the authors summarize the derivation of the Bayesian 2D affine ideal observer (details in Appendix). Let us first consider the case of only one 2D template. Assume that a template T and an input stimulus image S are represented as:

$$T=\begin{pmatrix} x_{\mathrm{T}}^1 & x_{\mathrm{T}}^2 & \ldots & x_{\mathrm{T}}^{\mathrm{n}} \\ y_{\mathrm{T}}^1 & y_{\mathrm{T}}^2 & \ldots & y_{\mathrm{T}}^{\mathrm{n}} \end{pmatrix}=\begin{pmatrix} X_{\mathrm{T}} \\ Y_{\mathrm{T}} \end{pmatrix}, S=\begin{pmatrix} x_{\mathrm{S}}^1 & x_{\mathrm{S}}^2 \\ y_{\mathrm{S}}^1 & y_{\mathrm{S}}^2 \\ \ldots & x_{\mathrm{S}}^{\mathrm{n}} \\ \ldots & y_{\mathrm{S}}^{\mathrm{n}} \end{pmatrix}=\begin{pmatrix} X_{\mathrm{S}} \\ Y_{\mathrm{S}} \end{pmatrix}. \quad (1)$$

A 2D affine transformation to the template T is

$$A\,T+T_{\mathrm{r}}=\begin{pmatrix} a & b \\ c & d \end{pmatrix} T+\begin{pmatrix} t_x & 0 \\ 0 & t_y \end{pmatrix}\begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \end{pmatrix}, \quad (2)$$

with $X^{\mathrm{T}} = \{a, b, c, d, t_x, t_y\} \in (-\infty, \infty)$. The authors assume that the stimulus image S is obtained by first applying a 2D affine transformation to the template T, then adding independent Gaussian noise $N(0, \sigma I_{2n})$, where I_{2n} is a $2n \times 2n$ identity matrix. Therefore the probability $P(S|T, X) = P(N = S - (A\ T + T_r))$ Hence,

$$P(S|T) = \int_{-\infty}^{\infty} P(S|T, X)\ P(X)\, dX \tag{3}$$

$$= \frac{1}{(2\pi\sigma^2)^{2n/2}} \int dX P(X) \exp\left(-\frac{\|S - AT - T_r\|^2}{2\sigma^2} \right) \tag{4}$$

where $P(X)$ is the prior probability distribution of X^{T}. Assume that

$$P(X) = \frac{1}{(2\pi\gamma^2)^3} \exp\left(-\frac{(X - X_0)^{\mathrm{T}}(X - X_0)}{2\gamma^2} \right),\ X_0^{\mathrm{T}} = (1,0,0,1,0,0), \tag{5}$$

which means that the prior probability distribution of a 2D affine transformation to a template is a Gaussian centered at the identity transformation. Given that the six variables

$$(a,b,c,d,t_x,t_y) = X^T \in (-\infty, \infty) \tag{6}$$

are independent of each other, the authors obtain the following by integration:

$$P(S|T) = \frac{1}{(2\pi\sigma^2)^{n-3}\gamma^6(n+\gamma^{-2})\det(Q')} \exp\left(\frac{\mathrm{var}(x_S) + \mathrm{var}(y_S) + \frac{\bar{x}^2 + \bar{y}^2}{n\gamma^2 + 1} - 2\sigma^2/n}{} \right) \tag{7}$$

$$\times \exp\left(-\frac{2\gamma^{-2} - tr(K^{*\mathrm{T}} Q (Q')^{-1} Q K^*)}{2\sigma^2} \right), \tag{8}$$

where

$$Q' \equiv Q + \gamma^{-2} I_2, \tag{9}$$

$$Q \equiv \begin{pmatrix} X_{\mathrm{T}} & \cdot & X_{\mathrm{T}} & X_{\mathrm{T}} & \cdot & Y_{\mathrm{T}} \\ Y_{\mathrm{T}} & \cdot & X_{\mathrm{T}} & Y_{\mathrm{T}} & \cdot & Y_{\mathrm{T}} \end{pmatrix}, \tag{10}$$

$$I_2 \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{11}$$

$$QK^* = QK + \gamma^{-2} I_2 \equiv (QK_1^* \quad QK_2^*) \tag{12}$$

$$QK = \begin{pmatrix} X_{\mathrm{T}} & \cdot & X_{\mathrm{S}} & X_{\mathrm{T}} & \cdot & Y_{\mathrm{S}} \\ Y_{\mathrm{T}} & \cdot & X_{\mathrm{S}} & Y_{\mathrm{T}} & \cdot & Y_{\mathrm{S}} \end{pmatrix}. \tag{13}$$

Note that under the assumption of the Gaussian prior probability distribution $P(X) = N(X_0, \gamma I_6)$, γ is the only free parameter. When $\gamma \to 0$, the prior becomes a δ-function and no transformation is allowed to the template T. Only the template T, not even its 2D rotations in the image plane are allowed in the matching process. Since this might be over-restrictive, the authors assume that the 2D rotations of the templates are automatically available to the ideal observer. Therefore, both γ and the number of 2D rotations of the template will be explored to search for the optimal performance. The authors also assume that the ideal observer knows the sequence of the feature points, but not which is the head and which the tail, so both possibilities will be considered.

2.2. 2D affine nearest neighbor model

In the above derivation, the prior probability of the 2D affine transformations is assumed to be Gaussian centered at each of the learned templates and their 2D rotations. Although the authors will search for the optimal performance with this prior, it is informative to know the ideal observer's performance when it only uses the 2D affine transformation that brings the stimulus and template to the closest possible match. In other words, the authors will consider the nearest neighbor solution for the problem, for which [11] have an analytic derivation.

Their model assumes that the stimulus and template are represented by 2D point features of known correspondence. The similarity measure between S and T is defined by the smallest Euclidean distance between the two $2 \times n$ matrices $(x_i, y_i)_{i=1}^{n}$ after both images are normalized to the same scale (a point that will be returned to). Image S can undergo an arbitrary 2D similarity transformation (rotation, translation and scaling) and image T an arbitrary 2D affine transformation. They showed that the smallest squared Euclidean distance D^2 between the two images is:

$$D^2(S, T) = 1 - \frac{tr(S + S \cdot T^{\mathrm{T}} T)}{\|T\|^2}, \tag{14}$$

where $tr[\cdot]$ is the trace of a matrix, $\mathbf{S}^+ = S^{\mathrm{T}}(SS^{\mathrm{T}})^{-1}$ is the pseudo-inverse of S and $\|T\|^2 = tr[TT^{\mathrm{T}}]$.

Only the Euclidean distance D, not the probability, is defined between two images in this nearest neighbor model. Thus, when there are multiple templates, either the summation of the D^2 themselves, or the summation of $\exp(-D^2/2\sigma^2)$ can be used for the similarity measure. The authors will use both in this paper and report the one that gives rise to the better performance.

2.3. GRBF model

The authors also simulate an improved version of the GRBF model originally presented in [8]. In [1], the model stored a set of 2D images $\{T_i\}$ of the prototype object. When a pair of stimulus images $\{S_1, S_2\}$ were presented, the model chose the image with a larger probability value from the following evaluation function:

$$\sum_i c_i \exp\left(-\frac{\|T_i - S\|^2}{2\sigma^2}\right), \quad (15)$$

where $\{c_i\}$ were obtained optimally when the learned templates themselves were used as input stimuli.

Although σ is the right number to use for the learned views, it is not necessarily the best choice for the novel views, since the model only approximates a novel view using weighted Gaussian summations from the learned views. In this paper, the authors will search for the optimal value of σ for the novel views by hand picking that gives rise to the best performance, for each individual object.

2.4. The 3D/2D polynomial model

An important theoretical question in object recognition is the amount of available information in images, from which the 3D structure of an object can be determined. This is the so called shape-from-views problem. The approach dates back to the classic work of the four-points-three-views theorem of structure-from-motion, in which [13] showed that three images of four non-coplanar labeled points under orthographic projection determine the 3D structure of the four points (with a depth reversion ambiguity). The authors now briefly review the state of the art of the shape-from-views problem before introducing the model that is closely related to the current study. To begin with, [13] further showed that if a fourth image is available, it can be verified as coming from the same object or not. While this assumes that the object structure is rigid, [14] showed that the rigidity of a labeled e-point non-planar structure can be verified from three images by checking a $6 \times n$ matrix. If the matrix has a full rank, then the structure is non-rigid, otherwise it is.

When only two images S and T are available, the authors can write a matrix (assuming no translation):

$$M = \begin{pmatrix} x_S^1 & x_S^2 & \cdots & x_S^n \\ y_S^1 & y_S^2 & \cdots & y_S^n \\ x_T^1 & x_T^2 & \cdots & x_T^n \\ y_T^1 & y_T^2 & \cdots & y_T^n \end{pmatrix}. \quad (16)$$

If rank $(M) = 3$, then S and T are from the same object and the object can undergo arbitrary 3D affine transformations. If rank $(M) > 3$, then the two images are not from the same object [15–19]. In the recognition scheme in [15], for example, the two stored images serve as the basis for recognition. When a third image is available, it can be verified as coming from the same object (the third image can be obtained by applying a certain affine transformation to the object before orthographic projection) or not.

Bennett et al. (1993) [12] have proposed a specific implementation for object verification with two images (one stored image T, one input image S). This is equivalent to checking whether the two images are consistent with an object that can undergo arbitrary 3D affine transformations. This model is appealing since its implementation is simple (only a polynomial calculation), it is specifically proposed as a candidate model for human object recognition and its proposed Gaussian noise model is closely related to the study in this paper.

It starts with four points in each image, one point is at the origin (0, 0) to handle the translation. The images S and T belong to the same (3D affine) object if and only if:

$$R = \text{determinant} \begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3} \\ d_{2,1} & d_{2,2} & d_{2,3} \\ d_{3,1} & d_{3,2} & d_{3,3} \end{pmatrix} = 0, \quad (17)$$

where $d_{i,j} = x_{i,S}x_{j,S} + y_{i,S}y_{j,S} - x_{i,T}x_{j,T} - y_{i,T}y_{j,T}$. Similar to the theorem in [11], this recognition polynomial does not specify how the polynomial changes its value when the feature points are perturbed with noise. Bennett et al. (1993) [12] suggest using R^2 as a measure of the goodness of fit between the two images. When an image has more than four points, the polynomial is divided into subsets of four points each and that the overall similarity is the summation: $R(x_1, \cdot)^2 + R(x_2, \cdot)^2 + \ldots$ When an object has multiple stored templates, the above summation will also be across these templates.

3. Experimental methods

The experimental paradigm is described in detail in [1]. The authors review its basics here. In a training phase, a subject first learned a 3D prototype wire object from 11 viewpoints under orthographic projection with monocular viewing. For the subsequent testing phase, two objects were created by adding independent 3D positional Gaussian noise to the vertices of the learned prototype. The variance of the noise added to one object, called the target, was fixed and that added to the other, the distractor, was always larger. The two objects were presented to the subject from the same viewpoint. The task in the testing phase was to pick the object that was more similar in shape and size to the learned prototype[3]. The distractor standard deviation was varied using a staircase procedure [20] in order to find the observer's threshold at the 75% correct. The smaller

[3] In order to define a proper probability measure so that an ideal observer can be provably optimal in the task, the authors define image similarity as the Euclidean distance between their vertex coordinates. Therefore, the more two images differ in size, the less similar they are. An ideal observer exploits this, therefore it is only fair to allow human subjects the same.

this threshold is, the better the performance will be. Two conditions were randomly intermixed from trial to trial: learned views—the two objects were presented from one of the 11 learned viewpoints; and novel views—the objects were presented from an arbitrary viewpoint in 3D rotation. The thresholds for these two conditions were tracked in parallel.

Four classes of objects were used. They were, in the increasing order of object regularity (Fig. 1): Balls—five balls randomly arranged in 3D; Irregular—the five balls were connected by four cylinders into a chain; Symmetric—the above irregular object were bilaterally symmetric; and V-Shaped—the two cylinders on each side of the above symmetric object were collinear, so the object itself became planar and symmetric. When independent noise was added to perturb the positions of the balls, the cylinders connecting them were adjusted accordingly. So the V-Shaped objects were no longer perfectly planar, symmetric and collinear, nor were the symmetric objects precisely symmetric. There were three objects in each class. Three naive subjects participated in the experiment.

The four models described above were given the same task as the subjects. Each object's image was represented by an ordered sequence of the (x, y) coordinates of the wire vertices. Only the direction of the ordered sequence was assumed unknown, which is equivalent to a reflection ambiguity in correspondence between the feature points. The standard deviation of the Gaussian noise added to the target object was assumed known by the four models. The authors simulated the discrimination performance of the four models, using the same objects as seen by the human observers.

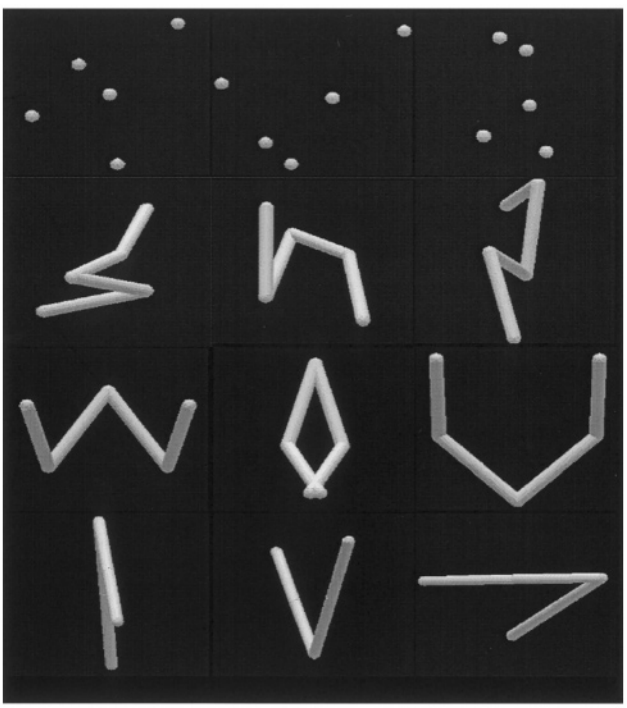

Fig. 1. Samples of the experimental stimuli. Top to bottom: Balls, Irregular, Symmetric and V-Shaped. (From [1], permitted by Vision Res.).

4. Predictions

For the Balls, Irregular and Symmetric objects, the authors expect that both human and the 2D affine ideal observer will perform better for the learned than for the novel views. In fact, for the learned views, the 2D affine ideal observer is the true ideal observer and human observers are necessarily less efficient due to internal noise. The question is, are they relatively more efficient for the novel views? In other words, are humans relatively better for the novel views than for the learned views as compared with the 2D affine ideal observer? A 'yes' answer implies that humans generalize from the learned to novel views better than the 2D affine ideal observer does. It further implies the 2D affine ideal observer cannot completely account for the human performance.

The authors are particularly interested in the Irregular and Symmetric objects and any differences between them. For the Balls objects, the authors expect that human subjects' performance will be poor. For the V-Shaped objects, the 2D affine ideal is the true ideal observer in the sense that it accurately models 3D viewpoint variations for planar objects (see footnote 2). These objects therefore serve as a control to verify that the ideal observer is doing the right thing.

If the performance of human observers relative to that of the 2D affine ideal observer (defined as the statistical efficiency [21]) is better for the novel views than for the learned views, then humans must have used a better recognition strategy than the 2D affine ideal. The reason is that the affine model only approximates the learned views, since the objects are not planar. If humans use a strategy of 2D affine transformations with independent 2D templates, then their performance relative to the 2D affine ideal for the novel views must be less than or equal to that for the learned views.

Due to the internal noise in the human visual system, the statistical efficiency for the learned views will be necessarily below 100%. Therefore, the statistical efficiency for the novel views may also be below 100%, even when it is greater than for the learned views. As long as the efficiency is higher for the novel views than for the learned views, the human observers have either employed a 2D transformation to the templates more complex than 2D affine transformations, or have not treated the templates as independent but effectively

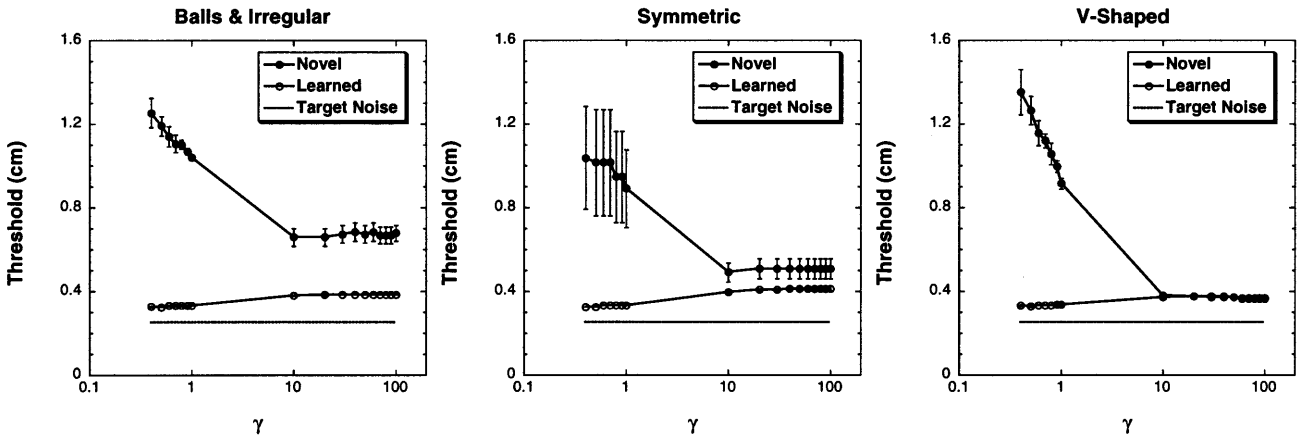

Fig. 2. Performance of the 2D affine ideal observer as a function of γ—the standard deviation of the Gaussian prior probability distribution, without additional rotational copies of any template ($m = 1$). The authors plot the Balls and Irregular objects together since they are treated the same by the model. The error bars are the standard errors (some are too small to be visible).

combined them to reconstruct the (partial) 3D structure of the object (view-combination, [15]).

The authors will employ a conservative (worst case) test for the 2D affine ideal and the GRBF model in the sense that the authors will select parameters that give rise to the best performance for the novel views. The models' performance for the learned views will be obtained with the parameters optimal for the novel views. (The parameters optimal for the novel views usually do not yield the best performance for the learned views.) In this way, the statistical efficiency for the novel views will be the lowest possible. This makes it more difficult to satisfy the hypothesis that the statistical efficiency for the novel views is higher than for the learned views. Consequently, if such a hypothesis is supported from the data, it will be evidence that the human observers use more 3D knowledge than implicit in the 2D affine transformations. The evidence will be strong in the sense that the best performance for the novel views is obtained by the experimenters, rather than by the models themselves. This is because it is difficult for the models to automatically search for the best performance when the two viewing conditions are randomly intermixed and no feedback is provided.

Finally, the polynomial model [12] predicts the same performance for the learned and novel views. This is because once the learned template is stored as the coefficients for the polynomial, the polynomial's mean value and variance are completely determined by the (x, y) coordinates of the feature points in the input image, learned and novel views alike. The variance associated with coding these (x, y) values can be assumed equal. Therefore, the model predicts that the human performance is viewpoint independent. It cannot account for human performance if human performance is different for the learned and novel views.

5. Results

5.1. The 2D affine ideal observer

Simulations were conducted to carry out the task for each of the 12 objects, learned and novel views respectively, with 2000 trials for each condition. These simulations were conducted for different γ values and different numbers of 2D rotated copies (m) of each template ($\gamma = 0.4$, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100; $m = 1$, 3, 5, 7, 9, 10, 11, 13, 15, 17, 19, 20, 40, 60, 80). It turns out that the smallest m ($= 1$) and sufficiently large γ values give rise to the best performance for the novel views (Figs. 2 and 3). The authors selected the model's best performance for the novel views (Balls: $m = 1$, $\gamma = 10$; Irregular and Symmetric: $m = 1$, $\gamma = 10$; V-Shaped: $m = 1, \gamma = 100$). Fig. 4 shows the statistical efficiency of the human observers relative to this 2D affine ideal observer (its derivation is in [1]). The authors conducted the Wilcoxon order test [22] for the 18 pairs of matched comparisons between the learned and novel views, for the Irregular and Symmetric objects (three objects each, three observers). The authors found that the efficiency for the novel views is statistically higher than for the learned views ($P < 0.02$; $T = 38$, $N = 18$, $z = 2.07$). This suggests that 2D affine transformations cannot account for the human performance.

There is a significant difference in statistical efficiency across the four types of objects ($F(3,6) = 18.25$, $P < 0.002$). Of particular interest is whether there is a difference between the Irregular and Symmetric objects. The efficiency for the Symmetric objects is higher than for the Irregular objects ($t(2) = 4.10$, $P < 0.03$). This suggests that the subjects may indeed have exploited symmetry in the task. This implies that subjects may take advantage of symmetry in 3D, since the 2D image of a novel view of a Symmetric object is almost always asymmetric.

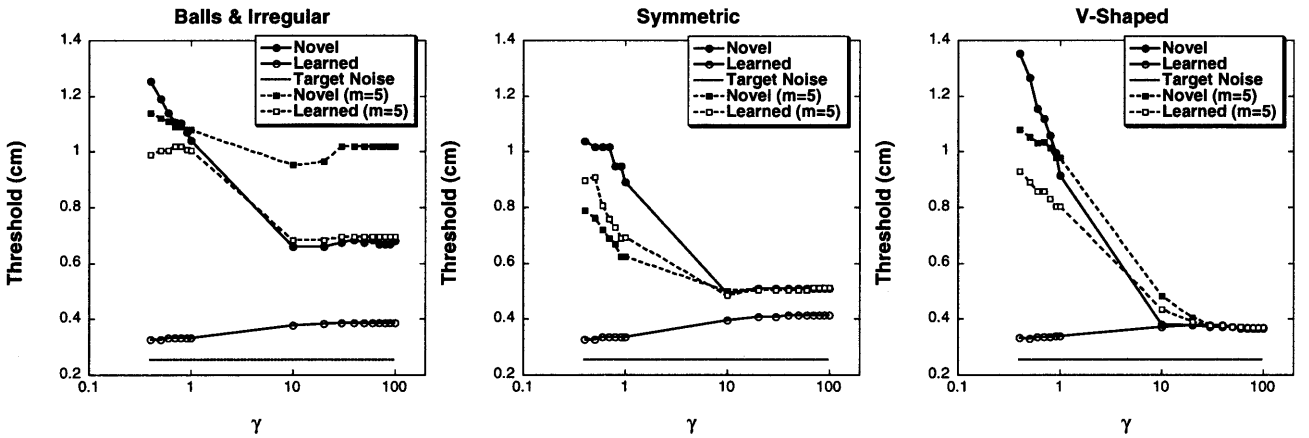

Fig. 3. Performance of the 2D affine ideal observer as a function of γ, for both $m = 1$ and 5 For clarity, the error bars are not shown.

In summary, the results that the human observers were more efficient for the novel than for the learned views and for the Symmetric than for the Irregular objects imply that the 2D affine ideal observer may not account for human performance. Three-dimensional structural information may have been exploited by the human visual system.

5.2. The remaining models

Fig. 5 shows the performance of the human observers, the 2D affine ideal observer, the 2D affine nearest neighbor model [11] (WW in short), the GRBF model [8] and the 3D/2D polynomial model [12]. The performance of the 2D affine nearest neighbor model was obtained by taking the summation of $\exp(-D^2(T_i, S)/2\sigma^2)$, rather than $D^2(T_i, S)$ directly[4]. The, suboptimal performance of the 2D affine nearest neighbor model is in part due to the fact that the model normalizes the size of each image first before computing the Euclidean distance. Thus the size information is not used at all, whereas in the studys task it is informative. The larger the noise is, the more likely the size is larger. This problem, however, does not apply to the rest of the models.

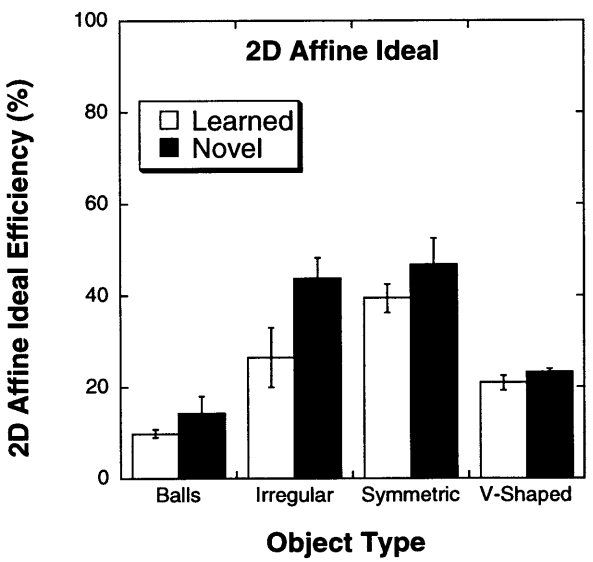

Fig. 4. Statistical efficiency of the human observers relative to the 2D affine ideal observer for the four types of objects. The error bars are standard errors between the three observers' scores. (Since the Wilcoxon analysis between the learned and novel view conditions is for matched pair comparison for each object and within each observer, the error bars cannot directly reflect the variance in the analysis. This applies to similar analysis below).

The authors make the following remarks. (1) The 3D/2D polynomial model's performance was very sensitive to the correspondence ambiguity, its threshold at least doubled when the correspondence is wrong. In contrast, the 2D affine model was much less sensitive to this ambiguity. This is because in the affine nearest neighbor model a normalization procedure is built in to align the two images by 2D linear transformations. In the 3D/2D polynomial model, however, no normalization procedure is available. When the correspondence is wrong, the model treats the input image as from a completely different object. This yields a poor polynomial evaluation and leads to a chance performance for the model at many instances, which was documented by the studys simulations. For this reason, the statistical efficiency will be plotted for the 2D affine nearest neighbor model with the correspondence ambiguity and the 3D/2D polynomial model with only the exact correspondence. (2) As expected, the 2D affine nearest neighbor model's threshold performance for the learned views was better than for the novel views, whereas the polynomial model's performance was about the same for both learned and novel views. (3) The 2D affine

[4] Using the metric of $\exp(-D^2(T_i, S)/2\sigma^2)$, the average thresholds for the first three object types were 0.46 and 0.90 cm, for the learned and novel views, respectively. They were 0.37 cm for the V-shaped object, for both learned and novel views. Using the $D^2(T_i, S)$ metric, they were 1.06, 1.07, 0.37 and 0.37 cm. So the first metric yields better performance.

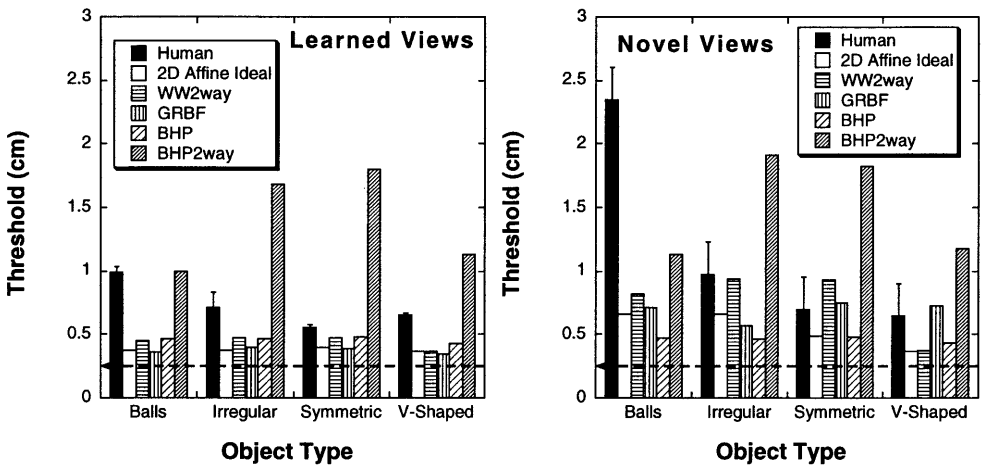

Fig. 5. Discrimination threshold of the human observers for the learned and novel views, for the 2D affine ideal observer, the [11] model with the two way correspondence ambiguity (WW2way), the GRBF model, the 3D/2D polynomial model [12] with the exact correspondence (BHP) and with the two way ambiguity correspondence (BHP2way). The threshold is defined as the standard deviation of the Gaussian noise added to the distractor object at 75% correct performance, for learned and novel views, respectively.

nearest neighbor model's performance for the V-Shaped objects was identical for the learned and novel views, as it should.

Fig. 6 shows human observers' statistical efficiency relative to the 2D affine nearest neighbor model. For the Irregular and Symmetric objects, the efficiency for the novel views is greater than for the learned views (Wilcoxon test, Irregular: $T=7$, $z=1.84$, $P<0.05$; Symmetric: $T=0$, $z=2.66$, $P<0.005$). This means that the 2D affine nearest neighbor matching cannot account for human data.

Fig. 6 also shows the human efficiency relative to the GRBF model, whose performance was obtained by hand picking the Gaussian variance that gives rise to the best performance for the novel views for each individual object. For all types of objects, the efficiency for the novel views was greater than for the learned views (Wilcoxon test, Balls: $T=8$, $z=1.72$, $P<0.05$; Irregular: $T=9$, $z=1.60$, $P<0.05$; Symmetric or V-Shaped: $T=0$, $z=2.66$, $P<0.005$). This means that the GRBF model, even when the standard deviation of its basis functions was allowed to (uniformly) vary to search for the best performance, still cannot account for the human performance.

Fig. 7 shows human observers' statistical efficiency relative to the polynomial model with exact correspondence. The absolute values of the efficiencies are high, but the overall pattern of the efficiency is similar to the 3D/2D (3D model/2D input) and 3D/3D (3D model/3D input) ideal observers (in [1], fig. 8, p. 561). The higher efficiencies for the learned views than for the novel views with the Balls, Irregular and Symmetric objects suggest that the 3D/2D polynomial recognition model, which predicts equal efficiencies, cannot account for human performance.

6. Discussion

The authors have derived an analytic formulation of a Bayesian model that gives rise to the best possible performance under 2D affine transformations. By using this model's performance as a benchmark for human performance, the authors have shown that the 2D affine ideal observer fails to account for human 3D object discrimination. Relative to this model, human statistical efficiency is higher for novel views than for learned views. If the statistical efficiencies had been 100% for both learned and novel views, the authors could have concluded with absolute certainty that the mechanisms used by human observers in this task is equivalent to a 2D affine observer. To what extent is it likely that a 2D affine observer (not ideal) could account for human performance? Excluding this possibility rests on at least five assumptions.

6.1. 2D observers for human 3D object recognition?

First, the conclusion that the observers did not use a 2D affine strategy, based on comparison of efficiencies between the novel and learned views, depends on the way in which internal noise in the visual system operates. For example, imagine that human observers are 2D affine ideal observer plus additive internal noise N. Then, according to Eqn. (E9) in [1], statistical efficiency E is the variance difference between the distractor and the target for the ideal observer Δs^{I} over that for the human observer $\Delta s^{\mathrm{H}}=\Delta s^{\mathrm{I}}+N$. It is reasonable to assume that N is the same for the novel views (n) and the learned views (l). Consequently,

$$E_{\mathrm{n}}=\frac{\Delta s_{\mathrm{n}}^{\mathrm{I}}}{\Delta s_{\mathrm{n}}^{\mathrm{I}}+N},\quad E_{l}=\frac{\Delta s_{\mathrm{l}}^{\mathrm{I}}}{\Delta s_{\mathrm{l}}^{\mathrm{I}}+N}. \tag{18}$$

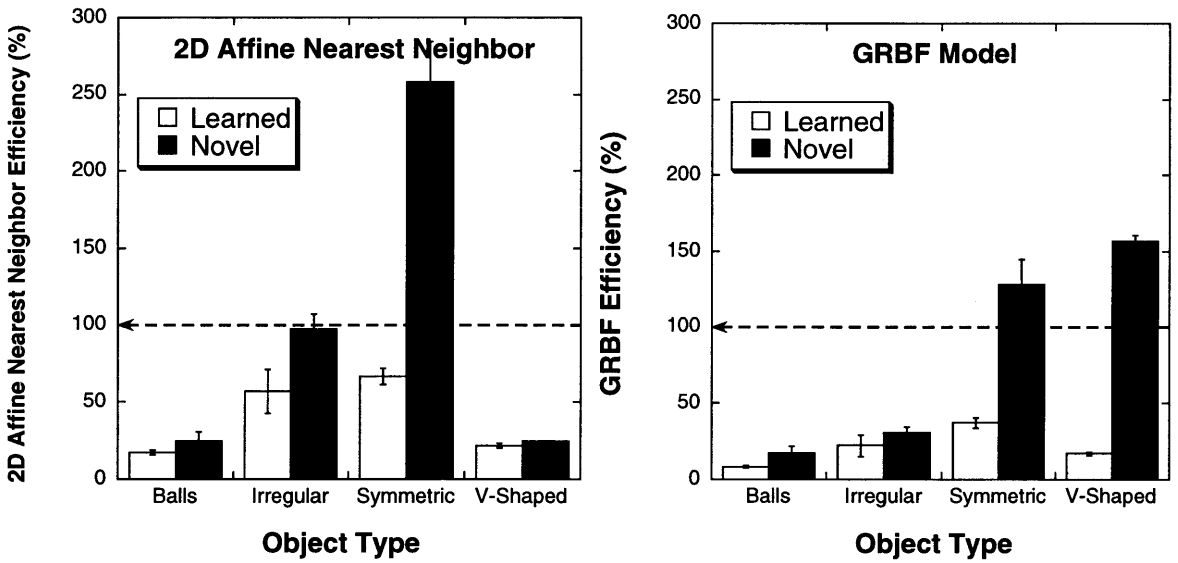

Fig. 6. Statistical efficiency of the human observers relative to the 2D affine nearest neighbor model and the GRBF model.

The authors have

$$\frac{E_n}{E_t}=\frac{1+\dfrac{N}{\Delta s_l^I}}{1+\dfrac{N}{\Delta s_n^I}}>1,(\Delta s_n^I>\Delta s_l^I>0). \tag{19}$$

This means that additive noise in itself increases the ratio of the efficiency for the novel views over the learned views. Although the equivalent internal noise thus derived is inconsistent between learned and novel views and between object types (Balls: learned views 0.75 cm^2, novel views 2.19 cm^2; Irregular: 0.22, 0.45; Symmetric: 0.14, 6.45), this only means that addition is unlikely the correct noise model. The authors cannot, however, exclude all possible ways in which internal noise increases the relative efficiency for the novel views.

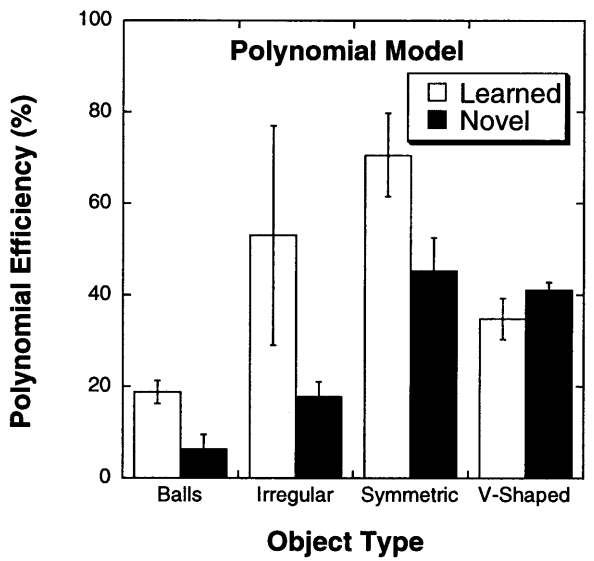

Fig. 7. Statistical efficiency of the human observers relative to the polynomial recognition model.

The second assumption is that each of the six variables in the affine transformation obeys a Gaussian prior probability distribution with the same variance. It is a problem because 2D affine transformation only approximates a 3D object rotation, therefore no 'correct' prior distributions ever exist. The choice is only a matter of convenience. On the other hand, however, the authors found that the performance of the 2D affine ideal observer stabilizes at optimal values so long as the variance of the Gaussian distributions are sufficiently large. It appears therefore that the specifics of the prior probability distribution are not critical.

The third assumption is that subjects did not learn the novel views during testing. If they did, these views could be used as additional 2D templates. Such learning would improve their performance for the novel views more than for the learned views. In contrast, the 2D affine ideal observer has only the fixed set of 2D templates. The authors considered this possibility in [1] by creating an additional 2D template for the 2D/2D ideal observer (with 2D rotations) after each test trial. The template was the average of the two test images and was a close approximation to a view of the prototype object since the two test images were from the same viewpoint. The incorporation of this learning improved the 2D/2D ideal observer's performance, but the efficiencies for the novel views of the Symmetric and V-Shaped objects were still above 100%. The authors did not simulate the 2D affine ideal observer with learning in this paper, but given the already small efficiency difference between the novel and learned views, the authors expect that the 2D affine ideal observer with learning could match human performance. In fact, however, when the authors tested the

third assumption, an analysis of the experimental data did not show learning by the human subjects [1].

The fourth assumption is that shaded images did not contribute an unfair advantage to the human observers. The human subjects had shading and occlusion information in addition to the vertex positions, whereas the models have available only the (x, y) vertex coordinates. The authors counted in [1] that the number of occlusion events was about the same for the learned and novel views and also argued that the shading information should be about the same for the learned and novel views. Therefore, it is unlikely that this additional image information alone should be responsible for any differential effect between the learned and novel views. One could argue that a differential effect can be obtained if subjects use only the vertex coordinates for the learned views and use everything possible for the novel views. The authors think that this scenario is possible but unlikely given that the learned and novel views in the human experiment were randomly intermixed. The authors are currently using silhouette images and thin wire objects to directly address this issue.

A fifth assumption is that the class of objects used in this study is representative of typical visual tasks that require fine shape discriminations. The objects were notably peculiar in their lack of substantial occlusions. It is true that previous studies used exactly the same type of objects to argue for a 2D template-based approach [8,4], therefore it is best to use the same objects to test the claim. But it remains a challenge to all computational studies to address everyday object recognition when occlusion is commonplace. On the other hand, the way in which an object is represented in the models, aside from the shading and partial occlusions, is not crucial for the results. So long as the authors perturb object vertex positions with Gaussian noise, the task is defined and the ideal observer performance is determined, no matter what representation is used. The choice of vertex (x, y) coordinates was a matter of mathematical convenience. If the authors represent the objects in terms of the lengths and relative angles of the cylinders, the authors would obtain the same ideal performance.

An additional point can be made from the symmetry condition. The authors noted that the efficiency is greater for the Symmetric than for the Irregular objects. Is this because the Symmetric objects are 'simpler,' in the sense that the viewing space is half as much for the Symmetric objects? This cannot explain why Symmetric objects have a greater efficiency, because the ideal observer's viewing space is also halved. In fact, the essence of ideal observer analysis and the measure of statistical efficiency is to take into account (or to normalize) any differences between different stimuli. Therefore, any efficiency difference reflects representation and processing differences in the brain, not in the stimulus. Therefore, the fact that the efficiency is higher for the Symmetric than for the Irregular objects indicates that subjects used 3D information of object symmetry in recognition.

Taken together, the results strongly suggest that 2D affine transformations are insufficient to account for the ability of humans to compensate for viewpoint changes in this task. What are the alternatives?

6.2. Is 3D structural information used for object recognition?

The fact that human statistical efficiency relative to the 2D affine ideal observer is greater for the novel than for the learned views, despite the best efforts to find the lowest efficiency possible for the novel views, indicates that human observers incorporate more knowledge of the regularities between views than that implicit in 2D affine transformations. The results also suggest that 2D affine nearest neighbor matching cannot account for the human performance. The fact that an 'ideal' prior probability distribution on all possible 2D affine transformations is unknown makes this result valuable in its own right. The authors can rank in increasing order of greater power and flexibility in approximating 3D novel object views from 2D template views, the 2D/2D ideal observer, the learning 2D/2D ideal observer, the Radial Basis Functions (RBF) model (in [1]) and in this paper the 2D affine nearest neighbor model, the GRBF model and the 2D affine ideal observer model. The results suggest that the human observers may use yet a more sophisticated strategy that incorporates knowledge of 3D structure, perhaps by means of view-combination [7].

Models of view-combination have either explicit or implicit knowledge that views arise from 3D object rotations. Three-dimensional constraints are built into the memory representations. By intelligently combining stored views, these models can, in principle, find nearly exact matches to novel views with orthographic projection [7].

Consider first an extreme and ideal case in which there is an explicit 3D model in memory. 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 plane and then using a measure of similarity to test for a satisfactory match with the 2D input (see for example [23]). Liu et al. (1995) [1] referred to the statistically optimal version of this model as a 3D/2D ideal observer. Despite its intuitive simplicity, a straightforward implementation of this scheme is in general not computationally feasible. An elegant solution (view-combination) to the computational difficulty was proposed by [15], who showed that as few as two views are sufficient to carry out the verification process

by checking the linear dependence of a third view on the two views. Recognition here assumes, albeit implicitly, that the object has 3D affine structure. In general, view-combination models exploit the inherent regularity in the collection of images resulting from a projection of an object.

The authors noted that that the statistical efficiency is greater for the Symmetric than for the Irregular objects implies that the human observers may have used 3D structural information, since 3D symmetry is inherently a 3D property. The authors cannot rule out, however, the possibility that 2D affine transformations account for substantial, though incomplete, portions of the human efficiency. This is illustrated by the fact that the statistical efficiencies for both the learned and novel views are below 100% and that their difference is no longer substantial, even though statistically significant.

It is important to note that ideal observer analysis is crucial to the conclusions. When object recognition performance falls off as an object rotates away from the learned views, it is difficult to distinguish whether the result can be accounted for by a view-approximation model [4], without an ideal observer analysis. The dependence of human performance on viewpoint might simply reflect the information for the task in the stimulus and not the specific functional constraints of the visual system. Until stimulus information is adequately accounted for, such a problem will remain unsolved. The ideal observer analysis makes it possible to distinguish these possibilities and suggests that view-approximation is not the whole story.

Acknowledgements

DK was supported by a grant from the National Science Foundation, contract number SBR-9631682. We thank Ronen Basri, David Jacobs, David Knill, Michael Langer, Pascal Mamassian, Bosco Tjan, Daphna Weinshall, the anonymous reviewers and in particular, John Oliensis, for many helpful discussions. Weinshall pointed out to us the Werman–Weinshall theorem. Part of this work was presented at the Hong Kong International Workshop on 'Theoretical Aspects of Neural Computation,' Hong Kong University of Science and Technology, 1997; European Conference on Visual Perception (ECVP), Helsinki, Finland, 1997; 'Neural Information Processing' (NIPS), Denver, Colorado, 1997; and 'International Conference on Computer Vision' (ICCV), Mumbai, India, 1998.

Appendix A. The 2D affine ideal observer

Without loss of generality, the authors consider the case of only one stored template. Assume that the template T and the input stimulus image S are represented as:

$$T=\begin{pmatrix} x_{\mathrm{T}}^1 & x_{\mathrm{T}}^2 & \cdots & x_{\mathrm{T}}^n \\ y_{\mathrm{T}}^1 & y_{\mathrm{T}}^2 & \cdots & x_{\mathrm{T}}^n \end{pmatrix}=\begin{pmatrix} \mathrm{X_T} \\ \mathrm{Y_T} \end{pmatrix}, \quad \mathrm{S} \tag{20}$$
$$=\begin{pmatrix} x_{\mathrm{S}}^1 & x_{\mathrm{S}}^2 & \cdots & x_{\mathrm{S}}^n \\ y_{\mathrm{S}}^1 & y_{\mathrm{S}}^2 & \cdots & y_{\mathrm{S}}^n \end{pmatrix}$$

A 2D affine transformation to the template T is

$$AT+T_{\mathrm{r}}=\begin{pmatrix} a & b \\ c & d \end{pmatrix}T+\begin{pmatrix} t_{\mathrm{x}} & 0 \\ 0 & t_{\mathrm{y}} \end{pmatrix}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \end{pmatrix} \tag{21}$$

with $X^{\mathrm{T}} \equiv (a,\ b,\ c,\ d,\ t_{\mathrm{x}},\ t_{\mathrm{y}}) \in (-\infty,\ \infty)$.

If the authors assume that the stimulus image S is obtained by first applying a 2D affine transformation to the template image T and then adding independent Gaussian noise $N(0,\ \sigma I_{2\mathrm{n}})$ to the resultant image, the authors have

$$P(S|T,A,T_{\mathrm{r}})=P(N=S-(AT+T_{\mathrm{r}})). \tag{22}$$

Let us calculate $S-(A\ T+T_{\mathrm{r}})$ first. Without loss of generality, the authors assume that the template image T is centered at the origin, i.e. $\Sigma_{i=1}^{\mathrm{n}}\, x_{\mathrm{T}}^{\mathrm{i}}=\Sigma_{i=1}^{\mathrm{n}}\, y_{\mathrm{T}}^{\mathrm{i}}=0$. The authors calculate the squared Euclidean distance of $\|S-(A\ T+T_{\mathrm{r}})\|^2$. More specifically, the squared Euclidean distance is

$$\|T_{\mathrm{x}}+aX_{\mathrm{T}}+bY_{\mathrm{T}}-X_{\mathrm{S}}\|^2+\|T_{\mathrm{y}}+cX_{\mathrm{T}}+dY_{\mathrm{T}}-Y_{\mathrm{S}}\|^2 \tag{23}$$

For the first term, given that $\Sigma_i x_{\mathrm{T}}^{\mathrm{i}}=\Sigma\ _i y_{\mathrm{T}}^{\mathrm{i}}=0$, the authors have

$$\|T_{\mathrm{x}}+aX_{\mathrm{T}}+bY_{\mathrm{T}}-X_{\mathrm{S}}\|^2$$
$$=\|T_{\mathrm{x}}-X_{\mathrm{S}}\|^2+\|aX_{\mathrm{T}}+bY_{\mathrm{T}}\|^2-2(aX_{\mathrm{T}}\cdot X_{\mathrm{S}}+bY_{\mathrm{T}}\cdot X_{\mathrm{S}}) \tag{24}$$

The first term on the right side of the Eq. (24) is

$$nt_{\mathrm{x}}^2-2t_{\mathrm{x}}\sum x_{\mathrm{S}}^{\mathrm{i}}+X_{\mathrm{S}}^2=n[(t_{\mathrm{x}}-\bar{x})^2+\mathrm{var}(x_{\mathrm{S}})],$$

where

$$\bar{x}=\frac{\sum x_{\mathrm{S}}^{\mathrm{i}}}{n},\ \mathrm{var}(x_{\mathrm{S}})=\frac{x_{\mathrm{S}}^2}{n}-(\bar{x})^2.$$

The last two terms in Eq. (24) is $a^2X_{\mathrm{T}}^2+b^2Y_{\mathrm{T}}^2+2abX_{\mathrm{T}}\cdot Y_{\mathrm{T}}-2(aX_{\mathrm{T}}\cdot X_{\mathrm{S}}+bY_{\mathrm{T}}\cdot X_{\mathrm{S}})$. So the total squared distance is

$$n[(t_x-\bar{x})^2+(t_y-\bar{y})^2+\mathrm{var}(x_S)+\mathrm{var}(y_S)]$$
$$+X_T^2(a^2+c^2)+Y_T^2(b^2+d^2) \tag{25}$$
$$+2(ab+cd)X_T\cdot Y_{\mathrm{T}}$$
$$-2(aX_{\mathrm{T}}\cdot X_{\mathrm{S}}+bY_{\mathrm{T}}\cdot X_{\mathrm{S}}+cX_{\mathrm{T}}\cdot Y_{\mathrm{S}}+dY_{\mathrm{T}}\cdot Y_{\mathrm{S}}). \tag{26}$$

The authors write

$$Q\equiv\begin{pmatrix} X_{\mathrm{T}} \\ Y_{\mathrm{T}} \end{pmatrix}(X_{\mathrm{T}}\quad Y_{\mathrm{T}})=\begin{pmatrix} X_{\mathrm{T}}\cdot X_{\mathrm{T}} & X_{\mathrm{T}}\cdot Y_{\mathrm{T}} \\ Y_{\mathrm{T}}\cdot X_{\mathrm{T}} & Y_{\mathrm{T}}\cdot Y_{\mathrm{T}} \end{pmatrix}, \tag{27}$$

$$QK = \begin{pmatrix} X_T \\ Y_T \end{pmatrix}(X_S \quad X_S) \equiv (QK_1 \quad QK_2). \tag{28}$$

Then the authors can write the squared distance as

$$n[(t_x - \bar{x})^2 + (t_y - \bar{y})^2 + \mathrm{var}(x_S) + \mathrm{var}(x_S)] \tag{29}$$

$$+ (a \quad b)Q\begin{pmatrix} a \\ b \end{pmatrix} - 2(a \quad b)QK_1 + (c \quad d)Q\begin{pmatrix} c \\ d \end{pmatrix}$$

$$- 2(c \quad d)QK_2. \tag{30}$$

Let

$$v_x \equiv \begin{pmatrix} a \\ b \end{pmatrix}, \quad v_y \equiv \begin{pmatrix} c \\ d \end{pmatrix}.$$

Completing the square e.g.

$$v_x^T Q v_x - 2v_x^T QK_1 \rightarrow (v_x - K_1)^T Q(v_x - K_1) - K_1^T QK_1$$

gives

$$n[(t_x - \bar{x})^2 + (t_y - \bar{y})^2 + \mathrm{var}(x_S) + \mathrm{var}(y_S)] + v_x'^T Q v_x'$$

$$- K_1^T QK_1 + v_y'^T Q v_y' - K_2^T QK_2,$$

where

$$v'_x \equiv v_x - K_1$$

and

$$v'_y \equiv v_y - K_2.$$

A.1. Uniform prior

Assume that $P(X) = C^{-1}$, where C is a normalization constant such that

$$1 = \int P(X)\, dX \tag{31}$$

this effectively assumes that $X^T \equiv (a, b, c, d, t_x, t_y)$ has a uniform distribution in R^6 and that C is necessarily infinite. So

$$P(S|T) = \int_{-\infty}^{\infty} P(S|X, T)P(X)\, dX \tag{32}$$

$$= \frac{1}{(2\pi\sigma^2)^n C}\int dX$$

$$\times \exp\left(- \frac{\|S - A(a,b,c,d)T - T_r(t_x,t_y)\|^2}{2\sigma^2} \right) \tag{33}$$

Let

$$t_x \rightarrow t'_x \equiv t_x - \bar{x}, \tag{34}$$

$$t_y \rightarrow t'_y \equiv t_y - \bar{y}, \tag{35}$$

$$a \rightarrow a' \equiv a - K_{11}, \tag{36}$$

$$b \rightarrow b' \equiv b - K_{21}, \tag{37}$$

$$c \rightarrow c' \equiv c - K_{12}, \tag{38}$$

$$d \rightarrow d' \equiv d - K_{22}. \tag{39}$$

The integral is:

$$P(S|T) = \frac{1}{C}\exp\left(- \frac{n(\mathrm{var}(x_S) + \mathrm{var}(y_S)) - tr(K^T QK)}{2\sigma^2} \right) \tag{40}$$

$$\frac{1}{(2\pi\sigma^2)^n}\int_{-\infty}^{\infty} da'db'dc'dd'dt'_x dt'_y$$

$$\times \exp\left(- \frac{nt_x'^2 + nt_y'^2 + v'_x Q v'_x + v'_y Q v'_y}{2\sigma^2} \right). \tag{41}$$

Now the authors use

$$\frac{1}{(2\pi\sigma^2)^{N/2}}\int_{-\infty}^{\infty} dx \exp\left(- \frac{x^T Mx}{2\sigma^2} \right) = (\det(M))^{-1/2},$$

where x is a length–**N** vector and **M** a symmetric $\mathbf{N} \times \mathbf{N}$ matrix. (Verify this by diagonalizing M and changing variables, then the integral just becomes a product of **N** independent Gaussian integrals.) Finally, the integral is

$$P(S|T) = \frac{1}{nC(2\pi\sigma^2)^{n-3}\det(Q)}$$

$$\exp\left(\frac{tr(K^T QK) - n(\mathrm{var}(x_S) + \mathrm{var}(y_S))}{2\sigma^2} \right). \tag{42}$$

A.2. Gaussian prior

Alternatively, the authors can assume that $X^T \equiv (a, b, c, d, t_x, t_y)$ obeys a Gaussian probability distribution

$$P(X) = \frac{1}{(2\pi\gamma^2)^3}\exp\left(- \frac{(X - X_0)^T(X - X_0)}{2\gamma^2} \right) \tag{43}$$

A reasonable assumption about X_0 is that this affine transformation is an identity transformation, with $X_0^T = (1, 0, 0, 1, 0, 0)$. The argument of the integral becomes proportional to

$$n[(t_x - \bar{x})^2 + (t_y - \bar{y})^2 + \mathrm{var}(x_S) + \mathrm{var}(y_S)] + \tag{44}$$

$$+ (a \quad b)Q\begin{pmatrix} a \\ b \end{pmatrix} - 2(a \quad b)QK_1 + (c \quad d)Q\begin{pmatrix} c \\ d \end{pmatrix}$$

$$- 2(c \quad d)QK_2 \tag{45}$$

$$+ \gamma^{-2}(t_x^2 + t_y^2 + (a-1)^2 + b^2 + c^2 + (d-1)^2) \tag{46}$$

$$= (n + \gamma^{-2})\left[\left(t_x - \frac{n\bar{x}}{n + \gamma^{-2}} \right)^2 + \left(t_y - \frac{n\bar{y}}{n + \gamma^{-2}} \right)^2 \right]$$

$$+ n(\bar{x}^2 + \bar{y}^2)\left(\frac{\gamma^{-2}}{n + \gamma^{-2}} \right) \tag{47}$$

$$+ n(\mathrm{var}(x_S) + \mathrm{var}(y_S)) \tag{48}$$

$$+ v_x^{*T} Q' v_x^* + v_y^{*T} Q' v_y^* - K_1^{*T} Q(Q')^{-1} QK_1^{*T}$$

$$- K_2^{*T} Q(Q')^{-1} QK_2^{*T} + 2\gamma^{-2}, \tag{49}$$

where

$$Q' \equiv Q + \gamma^{-2} I_2, QK^* = QK + \gamma^{-2} I_2 \equiv \begin{pmatrix} QK_1^* & QK_2^* \end{pmatrix},\ v^* = v - Q'^{-1} QK^*. \tag{50}$$

Similar arguments as before give

$$P(S|T) = \exp\left(- \frac{\mathrm{var}(x_S) + \mathrm{var}(y_S) + (\bar{x}^2 + \bar{y}^2)/(n\gamma^2 + 1)}{2\sigma^2/n} \right) \tag{51}$$

$$\times \exp\left(\frac{tr(K^T Q (Q')^{-1} QK) - 2\gamma^{-2}}{2\sigma^2} \right)$$

$$\times \frac{1}{(2\pi\gamma^2)^3} \int da'db'dc'dd'dt'_x dt'_y \tag{52}$$

$$\times \exp\left(- \frac{(n + \gamma^{-2})(t'^2_x + t'^2_y) + v'_x Q' v'_x + v'_y Q' v'_y}{2\sigma^2} \right) \tag{53}$$

$$P(S|T) = \frac{1}{(2\pi\sigma^2)^{n-3} \gamma^6 (n + \gamma^{-2}) \det(Q')}$$

$$\exp\left(- \frac{\mathrm{var}(x_S) + \mathrm{var}(y_S) + \dfrac{\bar{x}^2 + \bar{y}^2}{n\gamma^2 + 1}}{2\sigma^2/n} \right) \tag{54}$$

$$\times \exp\left(- \frac{2\gamma^{-2} - tr(K^{*T} Q (Q')^{-1} QK^*)}{2\sigma^2} \right). \tag{55}$$

As $\gamma \to \infty$ this goes to the former expression of Eq. (42).

References

[1] Liu Z, Knill DC, Kersten D. Object classification for human and ideal observers. Vision Res. 1995;35:549–68.

[2] Rock I, DiVita J. A case of viewer-cantered object perception. Cognit. Psychol. 1987;19:280–93.

[3] Tarr MJ, Pinker S. When does human object recognition use a viewer-cantered reference frame? Psychol. Sci. 1990;1:253–6.

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

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

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

[7] Ullman S. High-level Vision: Object Recognition and Visual Cognition. Cambridge, MA: MIT Press, 1996.

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

[9] Alter TD. Three-dimensional pose from three points using weak-perspective. IEEE Trans. Pattern Anal. Machine Intell. 1994;16:802–8.

[10] Basri R. Paraperspective ≡ affine. Int. J. Comput. Vis. 1996;19:169–79.

[11] Werman M, Weinshall D. Similarity and affine invariant distances between 2D point sets. IEEE Trans. Pattern Anal. Machine Intell. 1995;17:810–4.

[12] Bennett BM, Hoffman DD, Prakash C. Recognition polynomials. J. Opt. Soc. Am. A 1993;10:759–64.

[13] Ullman S. The interpretation of structure from motion. Proc. R. Soc. London B 1979;203:405–26.

[14] Weinshall D. Model-based invariants for 3D vision. Int. J. Comput. Vis. 1992;10:27–42.

[15] S. Ullman, R. Basri, Recognition by linear combinations of models, A.I. Laboratory Technical Report 1152, Massachusetts Institute of Technology, 1989.

[16] Ullman S, Basri R. Recognition by linear combinations of models. IEEE Trans. Pattern Anal. Machine Intell. 1991;13:992–1006.

[17] Bennett BM, Hoffman DD, Nicola JE. Structure from two orthographic views of rigid motion. J. Opt. Soc. Am. A 1989;6:1052–69.

[18] Huang TS, Lee CH. Motion and structure from orthographic projections. IEEE Trans. Pattern Anal. Machine Intell. 1989;2:536–40.

[19] C. Tomasi, T. Kanade, Shape and motion without depth, in: Proceedings of the Third International Conference on Computer Vision, Osaka, Japan, 1990.

[20] Watson AB, Pelli DG. QUEST: a Bayesian adaptive psychometric method. Percept. Psychophys. 1983;33:113–20.

[21] Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd, 1925.

[22] Hays WL. Statistics. New York: Holt, Rinehart and Winston, 1988.

[23] Basri R, Weinshall D. Distance metric between 3D models and 2D images for recognition and classification. IEEE Trans. Pattern Anal. Machine Intell. 1996;18:465–70.

[24] Liu Z. Viewpoint-dependency in object representation and recognition. Spat. Vis. 1996;9:491–521 Special Issue on Perceptual Learning and Adaptation in Man and Machine.

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