2D Affine Transformations Cannot Account for Human 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 observers' familiarity with objects' appearance. The more similar the objects are, the stronger this dependence will be, and the more important two-dimensional (2D) image information will be. The degree to which 3D structural information is used, however, 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 [3]. We now present results from models of generalized radial basis functions (GRBF), 2D nearest neighbor matching that allows 2D affine transformations, and a 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 template views, suggesting that humans generalize better to novel views from template views. The Bayesian estimator yields the optimal performance with 2D affine transformations and independent 2D templates. Therefore, no models of 2D affine operations with independent 2D templates account for the human performance.

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. But, in any object recognition study, the nature of the object representation has to be inferred from the recognition performance, by taking into account the contribution from the image information. When evaluating human performance, how can we separate the contributions of the image information from the representation? Ideal observer analysis provides a precise computational tool to answer this question. The ideal observer's recognition performance is restricted only by the available image information and is otherwise optimal, in the sense of statistical decision theory, irrespective of how the model is implemented. A comparison of human to ideal performance (in terms of *efficiency*) serves to normalize performance with respect to the image information for the task. We consider the problem of viewpoint dependence in human recognition.

A recent debate in human object recognition has focused on the dependence of recognition performance on viewpoint [1, 5]. Depending on the experimental conditions, an observer's ability to recognize a familiar object from novel viewpoints is impaired to varying degrees. A central assumption in the debate is the equivalence in viewpoint dependence between the representation in memory and recognition performance. In other words, the assumption is that a viewpoint dependent performance implies a viewpoint dependent representation, and that viewpoint independent performance implies 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 the performance is logically neither necessary nor sufficient for the viewpoint dependence of the representation. Image information has to be factored out first.

In addition to accounting for image information, the ideal observer has the additional virtue of being implementation free. Consider the GRBF model [4], as compared with human object recognition (see below). 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

$$\Sigma_i c_i \exp\left(-\frac{\|\mathbf{T}_i - \mathbf{S}\|^2}{2\sigma^2}\right), \qquad (1)$$

where c_i and σ are constants. The model's performance as a function of viewpoint parallels that of human observers. This observation has led to the conclusion that the human visual system may indeed, as does the model, use 2D stored views with GRBF interpolation to recognize 3D objects [2]. 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. There are typically too many possible models that can produce the same pattern of results.

In contrast, an ideal observer computes the optimal performance that is only limited by the stimulus information and the task. A constrained ideal is also limited by explicitly specified assumptions (e.g. a class of matching operations). It therefore yields the best possible performance among the class of models with the same stimulus input and assumptions. In this paper, we are particularly interested in constrained ideal observers that are restricted in functionally significant aspects (e.g., a 2D ideal observer that stores independent 2D templates and has access only to 2D affine transformations). The key idea is that a constrained ideal observer is the best in its class. So if humans outperform this ideal observer, they must have used more than what is available to the ideal. The conclusion that follows is strong: not only does the constrained ideal fail to account for human performance, but all implementations of it are also falsified as models of human recognition.

A crucial question in object recognition is the extent to which human observers model the geometric variation in images due to the projection of a 3D object onto a 2D image. At one extreme, we have shown that any model that compares the image to independent views (even if we allow for 2D rigid transformations of the input image) is insufficient to account for human performance [3]. At the other extreme, it is unlikely that variation is modeled in terms of rigid transformation of a 3D object template in memory. A possible intermediate solution is to match the input image to stored views, subject to 2D affine deformations. This is reasonable because, 2D affine transformations can capture a wider range of viewing conditions than 2D rigid transformations can.

In this study, we test whether any model limited to the independent comparison of 2D views, but with 2D affine flexibility, is sufficient to account for viewpoint dependence in human recognition. In the following section, we first define our experimental task, in which the computational models yield the provably best possible performance under their specified conditions. We then review the 2D ideal observer and GRBF model derived in [3], and the 2D affine nearest neighbor model in [6]. Our principal theoretical result is a closed-form solution of a Bayesian 2D affine ideal observer. We then compare human performance with the 2D affine ideal model, as well as the other three models. In particular, if humans can classify novel views of an object better than the 2D affine ideal, then our human observers must have used more information than that embodied by that ideal.

2 The observers

Let us first define the task. An observer looks at the 2D images of a 3D wire frame object from a number of viewpoints. These images will be called templates $\{\mathbf{T}_i\}$. Then two distorted copies 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 variance. The two objects are displayed from the same viewpoint in parallel projection, which is 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 original object. The observer's performance is measured by the variance (threshold) that gives rise to 75% correct performance.

Assume that the models are restricted to 2D transformations of the image, and cannot 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 us represent $\mathbf{S}$ and $\mathbf{T}_i$ by their (x, y) vertex coordinates: $\left(\mathbf{X} \quad \mathbf{Y} \right)^T$, where

$$\mathbf{X} = \left(x^1, x^2, \ldots, x^n\right), \mathbf{Y} = \left(y^1, y^2, \ldots, y^n\right). \quad (2)$$

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:

$$\mathbf{T}_i^r = \left(\mathbf{X}^r \quad \mathbf{Y}^r \right)^T, \quad (3)$$

$$\mathbf{X}^r = \left(x^n, \ldots, x^2, x^1\right), \mathbf{Y}^r = \left(y^n, \ldots, y^2, y^1\right). \quad (4)$$

We still denote the template set as $\{\mathbf{T}_i\}$. Therefore,

$$p(\mathbf{S}|\mathbf{O}) = \Sigma p(\mathbf{S}|\mathbf{T}_i) p(\mathbf{T}_i). \quad (5)$$

In what follows, we will compute $p(\mathbf{S}|\mathbf{T}_i)p(\mathbf{T}_i)$, with the assumption that

$$\mathbf{S} = \mathcal{F}\left(\mathbf{T}_i\right) + \mathbf{N}\left(\mathbf{0}, \sigma \mathbf{I}_{2n}\right), \quad (6)$$

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 nearest neighbor model, $\mathcal{F}$ represents the 2D affine transformation that minimizes $||\mathbf{S} - \mathbf{T}_i||^2$. For the 2D affine ideal observer, $\mathcal{F}$ represents all possible 2D affine transformations applicable to $\mathbf{T}_i$.

2.1 The 2D ideal observer

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 a template, the probability $p(\mathbf{S}|\mathbf{O})$ is an integration over all templates [3]:

$$\Sigma p(\mathbf{S}|\mathbf{T}_i)p(\mathbf{T}_i) \propto \Sigma \int d\phi \exp\left(-\frac{||\mathbf{S} - \mathbf{T}_i(\phi)||^2}{2\sigma^2}\right). \tag{7}$$

2.2 The GRBF model

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

$$\Sigma_i \int_0^{2\pi} d\phi c_i(\phi) N(||\mathbf{T}_j - \mathbf{T}_i(\phi)||, \sigma) = 1, j = 1, \ldots, \tag{8}$$

with which $\{c_i\}$ can be obtained optimally using singular value decomposition. When a pair of new stimuli $\{\mathbf{S}\}$ are presented, the optimal decision is to choose the one that is closer to the learned prototype, in other words, the one with a smaller value of

$$||1 - \Sigma \int_0^{2\pi} d\phi c_i(\phi) \exp\left(-\frac{||\mathbf{S} - \mathbf{T}_i(\phi)||^2}{2\sigma^2}\right)||. \tag{9}$$

2.3 The 2D affine nearest neighbor model

It has been proved in [6] that the smallest Euclidean distance $D(\mathbf{S}, \mathbf{T})$ between $\mathbf{S}$ and $\mathbf{T}$ is, when $\mathbf{T}$ is allowed a 2D affine transformation,

$$\mathbf{S} \to \frac{\mathbf{S}}{||\mathbf{S}||}, \mathbf{T} \to \frac{\mathbf{T}}{||\mathbf{T}||}, D^2(\mathbf{S}, \mathbf{T}) = 1 - \frac{tr(\mathbf{S}^+\mathbf{S} \cdot \mathbf{T}^T\mathbf{T})}{||\mathbf{T}||^2}, \tag{10}$$

where *tr* stands 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 $\Sigma \exp\left(-D^2(\mathbf{S}, \mathbf{T}_i)/2\sigma^2\right)$, or the smaller of $\Sigma D^2(\mathbf{S}, \mathbf{T}_i)$. (Both measures will be used and the results from the better one will be reported.)

2.4 The 2D affine ideal observer

We now calculate the Bayesian probability by assuming that the prior probability distribution of the 2D affine transformation, which is applied to the template $\mathbf{T}_i$,

$$\mathbf{AT} + \mathbf{T_r} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mathbf{T}_i + \begin{pmatrix} t_x & \cdots & t_x \\ t_y & \cdots & t_y \end{pmatrix}, \tag{11}$$

obeys a Gaussian distribution $\mathbf{N}(\mathbf{X_0}, \gamma\mathbf{I}_6)$, where $\mathbf{X}_0$ is the identity transformation

$$\mathbf{X}_0 = (a, b, c, d, t_x, t_y)^T = (1, 0, 0, 1, 0, 0)^T. \tag{12}$$

$$\Sigma p(\mathbf{S}|\mathbf{T}_i) = \Sigma \int dadbdcdd\, dt_x dt_y \tag{13}$$

$$\exp\left(-\frac{||\mathbf{AT}_i + \mathbf{T_r} - \mathbf{S}||^2}{2\sigma^2}\right) \tag{14}$$

$$= \Sigma \frac{\exp\left(\text{tr}\left(\mathbf{K}_i^T\mathbf{Q}_i\left(\mathbf{Q}_i'\right)^{-1}\mathbf{Q}_i\mathbf{K}_i\right)/2\sigma^2\right)}{C(n, \sigma, \gamma) det\left(\mathbf{Q}_i'\right)} \tag{15}$$

where $C(n, \sigma, \gamma)$ is a function of n, σ, γ;

$$\mathbf{Q}' = \mathbf{Q} + \gamma^{-2}\mathbf{I}_2, \mathbf{Q} = \begin{pmatrix} \mathbf{X_T} \cdot \mathbf{X_T} & \mathbf{X_T} \cdot \mathbf{Y_T} \\ \mathbf{Y_T} \cdot \mathbf{X_T} & \mathbf{Y_T} \cdot \mathbf{Y_T} \end{pmatrix}, \tag{16}$$

$$\mathbf{QK} = \begin{pmatrix} \mathbf{X_T} \cdot \mathbf{X_S} & \mathbf{Y_T} \cdot \mathbf{X_S} \\ \mathbf{X_T} \cdot \mathbf{Y_S} & \mathbf{Y_T} \cdot \mathbf{Y_S} \end{pmatrix} + \gamma^{-2}\mathbf{I}_2. \tag{17}$$

The free parameters are γ and the number of 2D rotated copies for each $\mathbf{T}_i$.

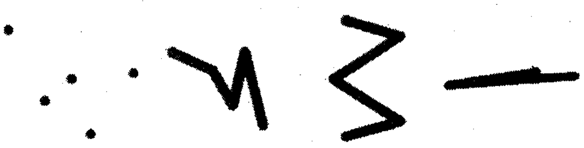

Figure 1: Stimulus classes: Balls, Irregular, Symmetric, and V-Shaped.

2.5 The human observers

Three naive subjects were tested with four classes of objects: Balls, Irregular, Symmetric, and V-Shaped (Fig. 1). There were three objects in each class. 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 orthographic projection, and viewed monocularly. 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. We predict that if the humans used a 2D affine strategy, then their performance *relative* to the 2D affine ideal observer should not be higher for the novel views than for the learned 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.). If human performance increases with the increasing structural regularity of the objects, this would lend support for the hypothesis that humans have used 3D information in the task.

2.6 Measuring performance

A stair-case procedure 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, and the result for the learned views was obtained accordingly. Likewise, for the 2D affine ideal, the number of 2D rotated copies of each template $\mathbf{T}_i$ and the value γ 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 test.

3 Results

Fig. 2 shows the threshold performance, i.e., the standard deviation of the Gaussian noise added to the distractor to maintain a 75% correct performance for the human observers and the models.

We use statistical efficiency to compare human to model performance. $\mathcal{E}$ is defined as the information used by humans relative to the ideal observer:

$$\mathcal{E} \equiv \left(\frac{d'_{human}}{d'_{ideal}}\right)^2 = \frac{\left(\sigma_d^{ideal}\right)^2 - (\sigma_t)^2}{\left(\sigma_d^{human}\right)^2 - (\sigma_t)^2}, \qquad (18)$$

where d' is the discrimination index, σ is the threshold — σ_t is that added to the target, and σ_d to the distractor [3]. Fig. 3 shows the statistical efficiency of the human observers relative to each of the four models.

We note in Fig. 3 that, relative to the affine observers, the efficiency for the novel views are higher than that for the learned views, 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 and novel views across the six objects and three subjects yielded a significant difference (binomial, $p < .05$). This suggests that the 2D affine ideal observer cannot account for the human observers' performance. We suggest therefore that 3D information was used by the human observers (e.g., 3D symmetry). This conclusion is supported in addition by the increasing efficiencies as the structural regularity increased from the Balls, Irregular, to Symmetric objects.

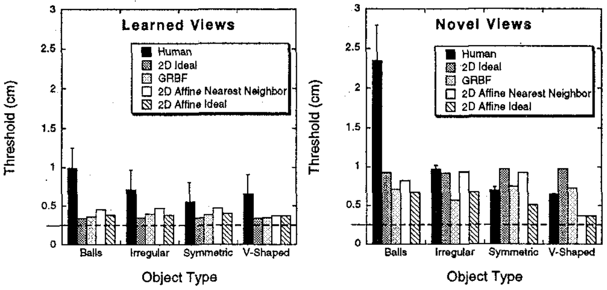

Figure 2: The thresholds for the learned and novel views, respectively. The dashed line is the standard deviation of the Gaussian noise added to the target.

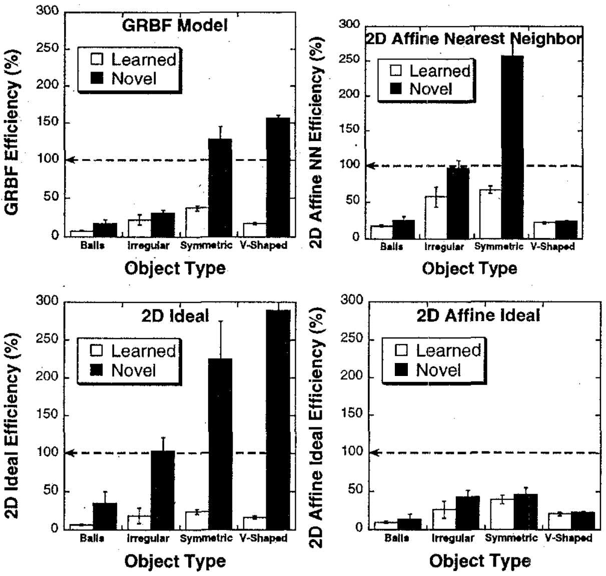

Figure 3: Statistical efficiencies of human observers relative to the four models.

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 humans, it is difficult to interpret which aspects of the model, if any, characterize 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. We showed here that models of independent 2D templates with 2D linear operations cannot account for the human performance, suggesting that our human observers may have used the templates to reconstruct a (crude) 3D structure of the object. This kind of strong conclusion rests on ideal observer analysis.

Appendix: 2D affine ideal observer

In this section, we derive the 2D affine ideal observer formulation. We consider the case of only one template. Assume that the template $\mathbf{T}$ and the input stimulus image $\mathbf{S}$ are represented as:

$$\mathbf{T} = \begin{pmatrix} x_{\mathbf{T}}^1 & x_{\mathbf{T}}^2 & \cdots & x_{\mathbf{T}}^n \\ y_{\mathbf{T}}^1 & y_{\mathbf{T}}^2 & \cdots & y_{\mathbf{T}}^n \end{pmatrix} = \begin{pmatrix} \mathbf{X_T} \\ \mathbf{Y_T} \end{pmatrix}, \quad (19)$$

$$\mathbf{S} = \begin{pmatrix} x_{\mathbf{S}}^1 & x_{\mathbf{S}}^2 & \cdots & x_{\mathbf{S}}^n \\ y_{\mathbf{S}}^1 & y_{\mathbf{S}}^2 & \cdots & y_{\mathbf{S}}^n \end{pmatrix} = \begin{pmatrix} \mathbf{X_S} \\ \mathbf{Y_S} \end{pmatrix}. \quad (20)$$

A 2D affine transformation to the template $\mathbf{T}$ is

$$\mathbf{A}\ \mathbf{T} + \mathbf{T_r} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mathbf{T} + \begin{pmatrix} t_x & \cdots & t_x \\ t_y & \cdots & t_y \end{pmatrix}, \quad (21)$$

with $\{a, b, c, d, t_x, t_y\} \in (-\infty, \infty)$.

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

$$p(\mathbf{S}|\mathbf{T}, \mathbf{A}, \mathbf{T_r}) = p(\mathbf{N} = \mathbf{S} - (\mathbf{A}\ \mathbf{T} + \mathbf{T_r})). \quad (22)$$

Let us calculate $\mathbf{S} - (\mathbf{A}\ \mathbf{T} + \mathbf{T_r})$ first. Without loss of generality, we assume that the template image $\mathbf{T}$ is centered at the origin, i.e.,

$$\Sigma_{i=1}^n x_{\mathbf{T}}^i = \Sigma_{i=1}^n y_{\mathbf{T}}^i = 0. \quad (23)$$

We now calculate the squared Euclidean distance of $||\mathbf{S} - (\mathbf{A}\ \mathbf{T} + \mathbf{T_r})||^2$. More explicitly, the squared Euclidean distance is

$$||\mathbf{T}_x + a\mathbf{X_T} + b\mathbf{Y_T} - \mathbf{X_S}||^2 + ||\mathbf{T}_y + c\mathbf{X_T} + d\mathbf{Y_T} - \mathbf{Y_S}||^2. \quad (24)$$

We now look at the first term, given Eqn. (23), we have

$$\begin{aligned} &||\mathbf{T}_x + a\mathbf{X_T} + b\mathbf{Y_T} - \mathbf{X_S}||^2 = && (25) \\ &\quad ||\mathbf{T}_x - \mathbf{X_S}||^2 + ||a\mathbf{X_T} + b\mathbf{Y_T}||^2 && (26) \\ &\quad -2\,(a\mathbf{X_T}\cdot\mathbf{X_S} + b\mathbf{Y_T}\cdot\mathbf{X_S}). && (27) \end{aligned}$$

The first term on the right side is

$$nt_x^2 - 2t_x\Sigma x_{\mathbf{S}}^i + \mathbf{X}_{\mathbf{S}}^2 = n[(t_x - \bar{x})^2 + var(x_{\mathbf{S}})], \quad (28)$$

$$\bar{x} = \Sigma x_{\mathbf{S}}^i/n, var(x_{\mathbf{S}}) = \mathbf{X}_{\mathbf{S}}^2/n - (\bar{x})^2. \quad (29)$$

The last two terms on the right side are

$$\begin{aligned} &a^2\mathbf{X}_{\mathbf{T}}^2 + b^2\mathbf{Y}_{\mathbf{T}}^2 + 2ab\mathbf{X_T}\cdot\mathbf{Y_T} - && (30) \\ &\quad 2\,(a\mathbf{X_T}\cdot\mathbf{X_S} + b\mathbf{Y_T}\cdot\mathbf{X_S}). && (31) \end{aligned}$$

So the total squared distance is

$$\begin{aligned} &n\left((t_x - \bar{x})^2 + (t_y - \bar{y})^2 + var\,(x_{\mathbf{S}} + y_{\mathbf{S}})\right) && (32) \\ &+\mathbf{X_T}^2\,(a^2 + c^2) + \mathbf{Y_T}^2\,(b^2 + d^2) && (33) \\ &+2(ab + cd)\mathbf{X_T}\cdot\mathbf{Y_T} && (34) \\ &-2\,(a\mathbf{X_T}\cdot\mathbf{X_S} + b\mathbf{Y_T}\cdot\mathbf{X_S}) && (35) \\ &+2\,(c\mathbf{X_T}\cdot\mathbf{Y_S} + d\mathbf{Y_T}\cdot\mathbf{Y_S}). && (36) \end{aligned}$$

We write

$$\begin{aligned} \mathbf{Q} &\equiv \begin{pmatrix} \mathbf{X_T} \\ \mathbf{Y_T} \end{pmatrix} \begin{pmatrix} \mathbf{X_T} & \mathbf{Y_T} \end{pmatrix} && (37) \\ &= \begin{pmatrix} \mathbf{X_T}\cdot\mathbf{X_T} & \mathbf{X_T}\cdot\mathbf{Y_T} \\ \mathbf{Y_T}\cdot\mathbf{X_T} & \mathbf{Y_T}\cdot\mathbf{Y_T} \end{pmatrix}, && (38) \end{aligned}$$

$$\mathbf{QK} = \begin{pmatrix} \mathbf{X_T} \\ \mathbf{Y_T} \end{pmatrix} \begin{pmatrix} \mathbf{X_S} & \mathbf{Y_S} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{QK_1} & \mathbf{QK_2} \end{pmatrix}. \quad (39)$$

Then we can write the squared distance as

$$\begin{aligned} &n[(t_x - \bar{x})^2 + (t_y - \bar{y})^2 + var\,(x_{\mathbf{S}} + y_{\mathbf{S}})] && (40) \\ &+\begin{pmatrix} a & b \end{pmatrix}\mathbf{Q}\begin{pmatrix} a \\ b \end{pmatrix} - 2\begin{pmatrix} a & b \end{pmatrix}\mathbf{QK_1} && (41) \\ &+\begin{pmatrix} c & d \end{pmatrix}\mathbf{Q}\begin{pmatrix} c \\ d \end{pmatrix} - 2\begin{pmatrix} c & d \end{pmatrix}\mathbf{QK_2}. && (42) \end{aligned}$$

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

Completing the square, e.g.,

$$\begin{aligned} &\mathbf{v_x}^T\mathbf{Q}\mathbf{v_x} - 2\mathbf{v_x}^T\mathbf{QK_1} \rightarrow && (43) \\ &\quad (\mathbf{v_x} - \mathbf{K_1})^T\,\mathbf{Q}\,(\mathbf{v_x} - \mathbf{K_1}) - \mathbf{K_1}^T\mathbf{QK_1}, && (44) \end{aligned}$$

gives

$$n\Sigma_{u=x,y}\left((t_u-\bar{u})^2+var\left(u_{\mathbf{S}}\right)\right) \tag{45}$$

$$+\mathbf{v}_{\mathbf{x}}'^{T}\mathbf{Q}\mathbf{v}_{\mathbf{x}}'-\mathbf{K_1}^{T}\mathbf{Q}\mathbf{K_1} \tag{46}$$

$$+\mathbf{v}_{\mathbf{y}}'^{T}\mathbf{Q}\mathbf{v}_{\mathbf{y}}'-\mathbf{K_2}^{T}\mathbf{Q}\mathbf{K_2}, \tag{47}$$

where $\mathbf{v}_{\mathbf{x}}' \equiv \mathbf{v_x}-\mathbf{K_1}$ and $\mathbf{v}_{\mathbf{y}}' \equiv \mathbf{v_y}-\mathbf{K_2}$.

Gaussian prior

We assume that

$$\mathbf{X}^T \equiv (a,b,c,d,t_x,t_y) \tag{48}$$

obeys a Gaussian probability distribution

$$p(abcd\,t_xt_y)=\frac{1}{(2\pi\gamma^2)^3}\exp\left(-\frac{(\mathbf{X}-\mathbf{X_0})^T(\mathbf{X}-\mathbf{X_0})}{2\gamma^2}\right) \tag{49}$$

A reasonable assumption about $\mathbf{X_0}$ is when the affine transformation is an identity transformation, with

$$\mathbf{X_0}^T=(1,0,0,1,0,0). \tag{50}$$

For simplicity, we use this $\mathbf{X_0}$ value from now on. The argument of the integral becomes proportional to

$$n[(t_x-\bar{x})^2+(t_y-\bar{y})^2+var\left(x_{\mathbf{S}}+y_{\mathbf{S}}\right)] \tag{51}$$

$$+\begin{pmatrix} a & b\end{pmatrix}\mathbf{Q}\begin{pmatrix} a \\ b\end{pmatrix}-2\begin{pmatrix} a & b\end{pmatrix}\mathbf{QK_1} \tag{52}$$

$$+\begin{pmatrix} c & d\end{pmatrix}\mathbf{Q}\begin{pmatrix} c \\ d\end{pmatrix}-2\begin{pmatrix} c & d\end{pmatrix}\mathbf{QK_2} \tag{53}$$

$$+\frac{t_x^2+t_y^2+(a-1)^2+b^2+c^2+(d-1)^2}{\gamma^2} \tag{54}$$

$$=\left(n+\gamma^{-2}\right)\left(t_x-\frac{n\bar{x}}{n+\gamma^{-2}}\right)^2 \tag{55}$$

$$+\left(n+\gamma^{-2}\right)\left(t_y-\frac{n\bar{y}}{n+\gamma^{-2}}\right)^2+2\gamma^{-2} \tag{56}$$

$$+n\left(\bar{x}^2+\bar{y}^2\right)\left(\frac{\gamma^{-2}}{n+\gamma^{-2}}\right) \tag{57}$$

$$+\mathbf{v}_{\mathbf{x}}^{*T}\mathbf{Q}'\mathbf{v}_{\mathbf{x}}^{*}+\mathbf{v}_{\mathbf{y}}^{*T}\mathbf{Q}'\mathbf{v}_{\mathbf{y}}^{*} \tag{58}$$

$$-\mathbf{K}_{\mathbf{1}}^{*T}\mathbf{Q}\left(\mathbf{Q}'\right)^{-1}\mathbf{QK}_{\mathbf{1}}^{*T}+n\,var\left(x_{\mathbf{S}}\right) \tag{59}$$

$$-\mathbf{K}_{\mathbf{2}}^{*T}\mathbf{Q}\left(\mathbf{Q}'\right)^{-1}\mathbf{QK}_{\mathbf{2}}^{*T}+n\,var\left(y_{\mathbf{S}}\right), \tag{60}$$

$$\mathbf{Q}'\equiv\mathbf{Q}+\gamma^{-2}\mathbf{I}_2, \mathbf{v}^*=\mathbf{v}-\mathbf{Q}'^{-1}\mathbf{QK}^*. \tag{61}$$

$$\mathbf{QK}^*=\mathbf{QK}+\gamma^{-2}\mathbf{I}_2\equiv\begin{pmatrix}\mathbf{QK}_{\mathbf{1}}^* & \mathbf{QK}_{\mathbf{2}}^*\end{pmatrix}, \tag{62}$$

$$p\left(\mathbf{S}|\mathbf{T}\right)=\exp\left(\frac{var\left(x_{\mathbf{S}}+y_{\mathbf{S}}\right)+\frac{\bar{x}^2+\bar{y}^2}{n\gamma^2+1}}{-2\sigma^2/n}\right) \tag{63}$$

$$\times\exp\left(\frac{\mathrm{tr}\left(\mathbf{K}^T\mathbf{Q}\left(\mathbf{Q}'\right)^{-1}\mathbf{QK}\right)-\frac{2}{\gamma^2}}{2\sigma^2}\right) \tag{64}$$

$$\times\frac{1}{(2\pi\gamma^2)^3}\int da'\ db'\ dc'\ dd'\ dt_x'\ dt_y' \tag{65}$$

$$\times\exp\left(-\frac{\left(n+\gamma^{-2}\right)\left(t_x'^2+t_y'^2\right)}{2\sigma^2}\right) \tag{66}$$

$$\times\exp\left(-\frac{\mathbf{v}_{\mathbf{x}}'\mathbf{Q}'\mathbf{v}_{\mathbf{x}}'+\mathbf{v}_{\mathbf{y}}'\mathbf{Q}'\mathbf{v}_{\mathbf{y}}'}{2\sigma^2}\right) \tag{67}$$

$$p(\mathbf{S}|\mathbf{T})=\frac{}{\left(2\pi\sigma^2\right)^{n-3}\gamma^6\left(n+\gamma^{-2}\right)\det\left(\mathbf{Q}'\right)} \tag{68}$$

$$\times\exp\left(-\frac{var\left(x_{\mathbf{S}}+y_{\mathbf{S}}\right)+\frac{\bar{x}^2+\bar{y}^2}{n\gamma^2+1}}{2\sigma^2/n}\right) \tag{69}$$

$$\times\exp\left(\frac{\mathrm{tr}\left(\mathbf{K}^{*T}\mathbf{Q}\left(\mathbf{Q}'\right)^{-1}\mathbf{QK}^*\right)-\frac{2}{\gamma^2}}{2\sigma^2}\right). \tag{70}$$

Acknowledgments

DK is supported by NSF grant SBR-9631682. We thank R Basri, D Jacobs, M Langer, B Tjan, D Weinshall, and in particular, J Oliensis, for helpful discussions.

References

[1] Biederman I & Gerhardstein G. Viewpoint dependent mechanisms in visual object recognition: a critical analysis. *J Exp Psych: HPP*, 21:1506–1514, 1995.

[2] Bülthoff H H & Edelman S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *PNAS*, 89:60–64, 1992.

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

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

[5] Tarr M & Bülthoff H H. Is human object recognition better described by geon-structural-descriptions or by multiple-views? *J Exp Psych: HPP*, 21:1494–1505, 1995.

[6] Werman M & Weinshall D. Similarity and affine invariant distances between 2D point sets. *PAMI*, 17:810–814, 1995.