Information-conserving object recognition

Margrit Betke *
Boston College
Chestnut Hill, MA 02167

Nicholas C. Makris †
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

Following the theory of statistical estimation, the problem of recognizing objects imaged in complex real-world scenes is examined from a parametric perspective. A scalar measure of an object's complexity, *which is invariant under affine transformation and changes in image noise level, is extracted from the object's Fisher information. The volume of Fisher information is shown to provide an overall statistical measure of the object's* recognizability *in a particular image, while the complexity provides an intrinsically physical measure that characterizes the object in any image. An* information-conserving *method is then developed for recognizing an object imaged in a complex scene. Here the term information-conserving means that the method uses all the measured data pertinent to the object's recognizability, attains the theoretical lower bound on estimation error for any unbiased estimate, and therefore is statistically optimal. This method is then successfully applied to finding objects imaged in thousands of complex real-world scenes.*

1 Introduction

Charge-coupled device (CCD) cameras typically produce scene images with extremely low but finite noise variance. In fact, for object recognition purposes in computer vision, an initial assumption often is that the noise can be altogether neglected so that the data at each pixel can be regarded as deterministic.

In the present investigation, however, we take an alternative approach that follows a strictly physical interpretation of classical estimation theory. First, we use experimental data to determine the joint probability distribution for the brightness measurements at each pixel in our CCD images. We use this to construct the likelihood function for any parameter set that is to be estimated given our image data. For our object recognition problem, we choose a parameter set that enables us to uniquely identify objects in complex real world scenes. It is significant that the form of the likelihood function in this physical approach is not at all arbitrary, but intrinsically depends upon the probability distribution of the brightness measurements no matter how low the corresponding noise variance is at each pixel, so long as it is finite. Moreover, it is the form of this likelihood function, not the level of the noise, that determines the optimal method for recognizing an imaged object.

To emphasize these issues, we show how a scalar measure of an object's *complexity*, which is invariant under affine transformation and changes in image noise level, can be extracted from the object's Fisher information. The volume of Fisher information is shown to provide an overall statistical measure of the object's *recognizability* in a particular image, while the complexity provides an intrinsically physical measure that characterizes the object in any image. We then derive a method for recognizing an object imaged in a complex scene that attains the theoretical lower bound on mean-square error for any unbiased estimate, and therefore is by definition *statistically optimal* and *information-conserving*. From the computer vision perspective, we consider the information-conserving property of this estimator to be most significant because it assures that the method uses all the measured data pertinent to the object's recognizability regardless of the noise level. Many popular edge-based methods, for example, discard a significant amount of information pertinent to an object's recognizability and are therefore inherently sub-optimal.

To illustrate our strictly physical approach with compelling examples, we focus attention in the present paper on the problem of recognizing objects that can be uniquely determined by the six parameters of an affine transformation as well as a seventh parameter that identifies the class of the object. Here, the affine transformation describes rigid body motion and lin-

*The work was supported by ONR grant N00014-95-1-0521. The author was with the Computer Vision Laboratory, University of Maryland, College Park, MD 20742. betke@cs.bc.edu.

†The author thanks ONR for support. The author was with the Naval Research Laboratory, Washington, DC 20375. Email: makris@keel.mit.edu.

ear distortions of a model object, while the class distinguishes it from other objects with the same affine parameters. For inherently three dimensional objects, the class must be supplemented by further parameterizations that account for such effects as variation in shading caused by changes in surface orientation with respect to a given source distribution and receiver geometry. For the recognition of flat objects in real world scenes, however, we show that such ancillary parameterizations are unnecessary so long as the object does not have a purely specular surface. This is because the optimal estimator for the affine parameters takes the form of a weighted filter that is invariant to the uniform variations in shading characteristic of such flat objects. This weighting is also necessary to discriminate against image ambiguities that are not explicitly accounted for in classical estimation theory. It is significant that these image ambiguities make the recognition problem inherently nonlinear. A global optimization procedure is therefore necessary to compute the filter output and obtain the optimal estimate.

The method's performance is evaluated experimentally by applying it to the problem of recognizing traffic signs in thousands of images of complicated outdoor scenes. In both our theoretical and experimental analysis, we find that recognizability is strongly dependent upon the object's complexity. We show how this measure becomes analogous to the complexity traditionally referred to in signal processing when the affine transformation is reduced to a 1-D shift in the position of a 1-D object.

2 The statistics of image brightness

Charge-coupled device (CCD) cameras do not output the intensity W of light, but instead a power-transformed intensity in 8-bit grey scales which we refer to as image *brightness* $I(x,y)$. The brightness is linearly proportional to $W^{-\gamma}(x,y)$ where a γ is a "gamma-correction," e.g., $\gamma = 2.2$. Experiments with the CCD video camera used in our vision system indicate that the standard deviation $\sigma(x,y)$ of the output $I(x,y)$ is not only small compared to the mean $m(x,y)$, but does not depend on the mean or on position (x,y) [2]. The noise, therefore, is additive and signal-independent, such that $\sigma(x,y) = \sigma$. We speculate that the noise is due to small mechanical vibrations between source and receiver, as well as electronic shot noise. Thermally induced fluctuations of natural light, however, are not a significant cause of errors in our measurements as is shown in Ref. [2].

Our measured average skew of -0.02 and kurtosis of 2.81 are so near to the corresponding Gaussian values of 0 and 3, respectively, that our data can be effectively modeled as Gaussian at each pixel By computation of the sample covariance of brightness between image pixels, our experiments also indicate that the brightness measurements are statistically independent across the pixels.

Let vector $\mathbf{I}$ represent image $I(x,y)$ where the rows of the image are appended into one column vector in lexicographic order. Each component I_k of vector $\mathbf{I}$ contains an independent intensity measurement $I(x,y)$ for $1 \leq k \leq MN$. Then the probability density for $\mathbf{I}$ is

$$P(\mathbf{I}) = \frac{1}{(2\pi\sigma^2)^{MN/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{MN}(I_k - m_k)^2\right). \quad (1)$$

3 Recognition as a parameter estimation problem

We use the six-dimensional vector $\mathbf{a} = (x_0, y_0, \theta_0, s_x, s_y, \alpha)$ to describe rigid body motion and linear distortions of an object q in an image with position $\mathbf{x_0} = (x_0, y_0)$, rotation θ_0, contractions s_x, s_y, and skew α which vanishes in a rectangular Cartesian coordinate system. For example, suppose the general Cartesian coordinates $\mathbf{x}' = (x', y')$ are related to the rectangular Cartesian system $\mathbf{x} = (x, y)$ by the 2-D affine transformation $\mathbf{x}' = \mathbf{A}\mathbf{x} - \mathbf{x_0}$, where

$$\mathbf{A} = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix} \begin{pmatrix} \cos\theta_0 & \sin\theta_0 \\ -\sin(\theta_0 + \alpha) & \cos(\theta_0 + \alpha) \end{pmatrix}. \quad (2)$$

A model object $q(x', y')$ in some ideal reference frame (x', y'), therefore, appears as a translated, rotated, contracted and skewed object $q(x, y; \mathbf{a})$ in the covariant reference frame (x, y) of an image. The parameters $\mathbf{a}$ are then measured within the image reference frame such that $-\infty < x_0, y_0 < \infty$, $0 \leq \theta_0 \leq 2\pi$, $-\pi/2 \leq \alpha \leq \pi/2$, and $0 < s_x, s_y < \infty$, where dilations occur for $0 < s_x, s_y < 1$ and contractions for $1 < s_x, s_y$.

To account for the possibility that distinct objects may have coincident vectors $\mathbf{a}$ we define an additional parameter ν that identifies the *class* of the object. For example, in traffic sign recognition, a slow sign is in a different class from a yield sign, although the two may have the same $\mathbf{a}$.

From the perspective of statistical estimation theory, recognizing an object is the same as estimating the parameters $\mathbf{a}$ and ν.

4 Parameter resolution: Fisher information, recognizability, and the coherence of objects in images

Let us consider the problem of recognizing an object of a given class in some scene. This can equivalently

be posed as the problem of estimating the parameter vector $\mathbf{a}$ given the image data $\mathbf{I}$. In this case, the likelihood function for $\mathbf{a}$ given the image data $\mathbf{I}$, is

$$P(\mathbf{I}|\mathbf{a}) = \frac{1}{(2\pi\sigma^2)^{MN/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{MN}(I_k - m_k(\mathbf{a}))^2\right) \tag{3}$$

where the mean $m_k(\mathbf{a})$ explicitly depends on the parameters to be estimated. The form of this likelihood function, given our CCD data, is very different from that in active radar, laser, and sonar imaging where nonlinear speckle noise is found [6].

The lower bound on the mean-square error in any unbiased estimate $\hat{\mathbf{a}}$ can be expressed as $\mathrm{E}[(\hat{\mathbf{a}}-\mathbf{a})(\hat{\mathbf{a}}-\mathbf{a})^T] \geq \mathbf{J}^{-1}$, where the Fisher information matrix $\mathbf{J}$ is defined by

$$J_{ij} = -\mathrm{E}\left[\frac{\partial^2}{\partial a_i \partial a_j} \ln P(\mathbf{I}|\mathbf{a})\right] \tag{4}$$

$$= \frac{1}{\sigma^2}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\left(\frac{\partial m(x,y;\mathbf{a})}{\partial a_i}\frac{\partial m(x,y;\mathbf{a})}{\partial a_j}\right). \tag{5}$$

Here the image mean $m(x,y;\mathbf{a})$ only depends on the parameter vector $\mathbf{a}$ for those pixels $(x,y) \in O^+$ that constitute the expected object $q(x,y;\mathbf{a})$ and any neighboring pixels that are affected by small changes in $\mathbf{a}$. The Fisher information matrix, therefore, can be reduced to

$$J_{ij} = \frac{1}{\sigma^2}\sum_{(x,y)\in O^+}\frac{\partial q(x,y;\mathbf{a})}{\partial a_i}\frac{\partial q(x,y;\mathbf{a})}{\partial a_j}. \tag{6}$$

It is significant that any of the diagonal entries of the bound can be expressed as

$$\mathrm{E}[(\hat{a}_i - a_i)^2] \geq [\mathbf{J}^{-1}]_{ii} = \frac{\sigma^2}{E}\ell_i^2, \tag{7}$$

where the object *energy* $E = \sum_{(x,y)\in O}|q(x,y;\mathbf{a})|^2$ and the *coherence scale*

$$\ell_i = \left([\mathbf{J}^{-1}]_{ii}\,\frac{E}{\sigma^2}\right)^{\frac{1}{2}} \tag{8}$$

for parameter a_i are physical descriptors of the object which are only invariant under rigid body motion. The coherence scale ℓ_i measures the sensitivity of the object to variations in parameter a_i and, therefore, can be interpreted as the width of the object's autocorrelation peak over lags in a_i. An object with relatively high sensitivity to parameter a_i, for example, will have a relatively narrow autocorrelation peak. The error in estimating parameter a_i, therefore, increases with the corresponding object coherence scale ℓ_i and additive noise variance, but decreases with object energy.

We define the *coherence volume* V to be the scalar measure characterizing the combined n_a-dimensional variations of the object, where n_a is the length of $\mathbf{a}$,

$$V = \left(\frac{E}{\sigma^2}\right)^{\frac{n_a}{2}}|\mathbf{J}|^{-\frac{1}{2}}, \tag{9}$$

where $|\mathbf{J}|$ is the determinant of the Fisher information matrix. The lower bound can then be written as

$$\mathbf{J}^{-1} = \mathbf{J}_{adj}\left(\frac{\sigma^2}{E}\right)^{n_a}V^2, \tag{10}$$

where $\mathbf{J}_{adj}$ is the adjugate matrix of $\mathbf{J}$ [8]. These coherence scales have compelling physical meanings. We consider the interpretation of $\mathbf{J}$ as an information measure to be far more useful than its interpretation as the inverse of the theoretical lower bound on estimation error. For example, in the type of optical pattern recognition problems encountered with low variance CCD camera measurements, the associated bounds on object positional resolution fall in the sub-pixel regime, and are somewhat of an overkill. On the other hand, because the volume $|\mathbf{J}|$ of Fisher information is inversely proportional to the limiting mean-square resolutional volume of the parameters that uniquely specify the object, we consider it to be a scalar measure of the object's *recognizability* in a given image. By Eq. (10) it is seen that there is a direct relationship between this *recognizability* measure and the *physical components* of the Fisher information, namely, the object's *coherence volume* and *energy*. For example, within a given image, where the additive noise variance is uniform, the information volume $|\mathbf{J}|$ only varies with the object's *coherence volume* and *energy*. The noise variance, therefore, factors out under variations in object recognizability, regardless the noise level. This shows that it is the physical structure of the likelihood function and not the level of the noise that is most important in properly formulating the recognition problem.

4.1 Position resolution

We first derive the lower bound on the error for any unbiased position estimate of an object with known rotation, contraction and skew. Given the true position $(a_1, a_2) = (x_0, y_0)$, the Fisher information matrix, with elements $J_{ij} =$

$$\frac{1}{\sigma^2}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\left(\frac{\partial q(x-x_0,y-y_0)}{\partial a_i}\frac{\partial q(x-x_0,y-y_0)}{\partial a_j}\right), \tag{11}$$

can be expressed by a spatial "bandwidth matrix" $\mathbf{B} = \sigma^2/E\mathbf{J}$ that characterizes the object. To do so, it

is convenient to let the double sum in Eq. (11) be replaced by a continuous double-integral so that $q(x,y)$ and $Q(u,v)$ can be defined as Fourier transform pairs

$$Q(u,v) = \iint_O q(x,y)e^{-j2\pi(xu+yv)}dxdy \quad \text{and}$$
$$q(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Q(u,v)e^{j2\pi(xu+yv)}dudv$$

where $(\Delta x)^2 = dxdy$ is the pixel area. The four elements of $\mathbf{B}$ can then be defined by a mean-square bandwidth B_x^2 in x,

$$B_x^2 = \frac{(2\pi)^2}{(\Delta x)^2 E}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2|Q(u,v)|^2dudv, \quad (12)$$

a mean-square bandwidth B_y^2 in y,

$$B_y^2 = \frac{(2\pi)^2}{(\Delta x)^2 E}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v^2|Q(u,v)|^2dudv, \quad (13)$$

and a cross-term

$$B_{xy}^2 = B_{yx}^2 = \frac{(2\pi)^2}{(\Delta x)^2 E}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uv|Q(u,v)|^2dudv, \quad (14)$$

with the aid of Parseval's Theorem $(\Delta x)^2 E = \iint_O |q(x,y)|^2dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|Q(u,v)|^2dudv$. These definitions for the object's mean-square spatial bandwidth are similar to those introduced for onedimensional signal waveforms by Gabor [4]. A distinction lies in the positive-semidefinite nature of our object brightness data versus the zero-mean nature of modulated signal waveform data. As a result, our mean-square bandwidths are defined about zero spatial frequency, as in Ref. [6], while those in the signal processing literature are defined about some average frequency that approximates the carrier frequency for narrowband signals.

Given these definitions and the derivative rule for Fourier transform pairs, the lower bound on position recognition can be expressed as

$$\mathbf{J}^{-1} = \frac{\sigma^2}{E}\mathbf{B}^{-1} = \frac{\sigma^2}{E}\begin{pmatrix} B_y^2 & -B_{xy}^2 \\ -B_{xy}^2 & B_x^2 \end{pmatrix} A_{x_0,y_0}^2, \quad (15)$$

where

$$A_{x_0,y_0} = |\mathbf{B}|^{-\frac{1}{2}} \quad (16)$$

is the *coherence area* of the object, which follows from Eq. 9, where $V = A_{x_0,y_0}$ for this 2-D scenario. For example, the lower bound for estimating x_0 is simply

$$\mathrm{E}[(\hat{x}_0 - x_0)^2] \geq J_{x_0}^{-1} = \frac{\sigma^2}{E}\ell_{x_0}^2, \quad (17)$$

where coherence length scale ℓ_{x_0} equals $B_y^2/|\mathbf{B}|$ or $B_y^2 A_{x_0,y_0}^2$, and the lower bound for y_0 is

$$\mathrm{E}[(\hat{y}_0 - y_0)^2] \geq J_{y_0}^{-1} = \frac{\sigma^2}{E}\ell_{y_0}^2, \quad (18)$$

where ℓ_{y_0} equals $B_x^2 A_{x_0,y_0}^2$. This analysis provides a 2-D extension of the well-known relationship between a 1-D signal's mean-square bandwidth and the optimal resolution attainable in an estimate of its position [3]. While the coherence length scales ℓ_{x_0} and ℓ_{y_0} could have been obtained directly from Eq. 8 without introducing the mean-square bandwidth concept, this would have circumvented both the historical perspective and an important physical interpretation.

The coherence areas and coherence length scales of two traffic signs are compared in Fig. 1. The shown stop sign has a much smaller coherence area than the shown European no-entry sign. Its position, therefore, can be resolved much more easily than the European no-entry sign's position.

The bound on position estimation error is not invariant to changes in object rotation, as is shown in Ref. [2] by principal component analysis.

4.2 Angular resolution

Assume only the rotation θ_0 of the object about some point in the image plane is unknown. By Eq. 8, the angular coherence scale for object rotation is

$$\ell_{\theta_0} = \left(\frac{E}{\sum_{(x,y)\in O^+}\left|\frac{\partial q(x,y)}{\partial\theta_0}\right|^2}\right)^{\frac{1}{2}}. \quad (19)$$

This leads to the bound

$$\mathrm{E}[(\hat{\theta}_0 - \theta_0)^2] \geq J_{\theta_0}^{-1} = \frac{\sigma^2}{E}\ell_{\theta_0}^2 \quad (20)$$

on angular resolution of the object, which is invariant to changes in object position, since $\frac{\partial x_0}{\partial\theta_0}$ and $\frac{\partial y_0}{\partial\theta_0}$ vanish, but depends on contraction and shear of the object, as shown in Ref. [2].

4.3 Contractional resolution

Finally, assume that only the object's contractional distortions s_x and s_y are unknown. Then, for 2-D parameter vector $(a_1, a_2) = (s_x, s_y)$, where $s_x, s_y > 0$, $\mathbf{J}$ is a 2×2 matrix with elements defined in Eq. 6. The coherence area A_{s_x,s_y} and coherence length scales ℓ_{s_x}, ℓ_{s_y} are then dependent, by Eq. 9, on both diagonal and cross terms of the Fisher information matrix, such that

$$A_{s_x,s_y} = \left(\frac{E}{\sigma^2}\right)|\mathbf{J}|^{-\frac{1}{2}}, \quad (21)$$

$\ell_{s_x} = \left([\mathbf{J}^{-1}]_{11}\frac{E}{\sigma^2}\right)^{\frac{1}{2}}$ and $\ell_{s_y} = \left([\mathbf{J}^{-1}]_{22}\frac{E}{\sigma^2}\right)^{\frac{1}{2}}$. The bounds for contractional resolution are then $\mathrm{E}[(\hat{s}_x - s_x)^2] \geq \frac{\sigma^2}{E}\ell_{s_x}^2$, and $\mathrm{E}[(\hat{s}_y - s_y)^2] \geq \frac{\sigma^2}{E}\ell_{s_y}^2$. See Fig. 2.

5 The complexity of imaged objects

According to standard usage, an object is considered to be *complex* if it is "composed of elaborately interconnected parts." We may gather from this that

as *complexity* increases so does the number of interconnected parts. These ideas help formulate a quantitative definition for the complexity of an imaged object. Let us first consider two objects of exactly the same dimensions but of different complexity that are imaged in an otherwise empty scene. For example, let the more complex object be a grey-scale Mona Lisa without a picture frame, the less complex object be a blank white canvas of the same dimensions, and the empty background be solid black. By their like dimensions, the two objects occupy the same overall area. As may be inferred from their descriptions, however, the two objects have vastly differing coherence areas. Let us consider a coherence area small if the ratio of it to the overall object area is much less than one. Then the Mona Lisa's coherence area will be small due to its large number of "elaborately interconnected parts," but the number of coherence areas or cells that fit into the Mona Lisa's overall area will be large. Conversely, the coherence area of the blank canvass will not be small, but the number of coherence cells that fit into the blank canvass' overall area will be near unity. We may consider the overall object area as an *outer scale* and the coherence area as an *inner scale* for variations in an object's 2-D position. It is the ratio of such an outer scale to an inner scale that determines the number of coherence cells in the object, also referred to as its degrees of freedom, which can be interpreted as its gain in sensitivity under transformation over the empty object space. By the foregoing argument, this ratio also serves as a quantitative measure of an object's complexity.

Generalizing these concepts, we define the *outer volume* under affine transformation, denoted by S, to be the product of the outer scales for 2-D positional transformation, rotation, 2-D contractions, and skew. These are, respectively, the object area A, 2π, unity, and π. The complexity of an object under affine transformation is then the ratio of this outer volume to the coherence volume V defined in Eq. 9, so that

$$C = \frac{S}{V} = A 2\pi^2 \left(\frac{\sigma^2}{E}\right)^{\frac{n_d}{2}} |\mathbf{J}|^{\frac{1}{2}}. \qquad (22)$$

The complexities of various traffic signs are compared in Fig. 3.

When the affine transformation is reduced to a 2-D translation, the relevant *positional complexity* becomes $C_{x_0,y_0} = \frac{A}{A_{x_0,y_0}}$, where the coherence area A_{x_0,y_0} is given in Eq. 16. When the translation is restricted to 1-D, the above complexity becomes analogous to that used in the signal processing literature for the analysis of complex waveforms [3]. Similarly, we define the *rotational complexity* of an object by $C_{\theta_0} = \frac{2\pi}{\ell_{\theta_0}}$, and the *contractional complexity* by $C_s = \frac{1}{A_{s_x,s_y}}$, where the rotational coherence scale ℓ_{θ_0} is defined in Eq. 19 and the contractional coherence area A_{s_x,s_y} in Eq. 21. The rotational and contractional complexities of the traffic sign models, charted in Fig. 4, are consistent with qualitative appraisals of the inherent rotational and contractional symmetries of the signs.

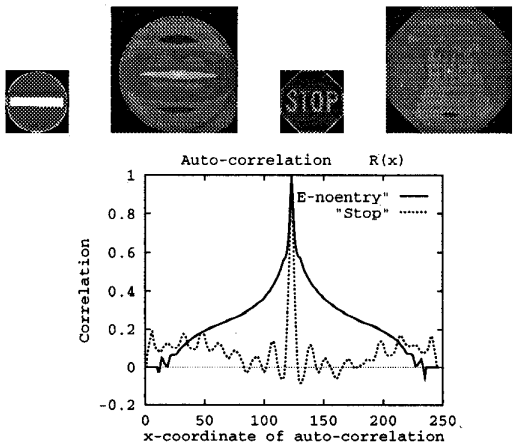

Figure 1: Two traffic signs and their 2D-autocorrelation surfaces. The white centers of the autocorrelation surfaces correspond to the coherence areas of the signs. European no-entry sign's coherence area of 2.2 % of the sign's area is much larger than the stop sign's, which is 0.4 %. This indicates that the position of the stop sign can be resolved more easily than the position of the European no-entry sign. Below are 1D-horizontal slices through the center of the signs' autocorrelation surfaces, where y-positions are fixed and x-positions vary. The stop sign's horizontal position can be resolved better than the European no-entry sign's because of its narrower autocorrelation peak-width and shorter coherence length.

6 Image edges

There is an important connection between the positional Fisher information of an object and "edge-based" recognition. Both require computation of the spatial gradient $(\frac{\partial q(x,y)}{\partial x}, \frac{\partial q(x,y)}{\partial x})$ of the expected object. By Eq. 11, however, the positional Fisher information integrates gradient factors over the entire object. This includes both slowly varying brightness contributions over the entire area of the object as well as rapid variations at the object's edges. A priori there is no way to judge which of these will make the dominant contribution to the Fisher information. In spite of this basic fact, edge-based recognition methods threshold the gradient magnitude over the object so as to discard

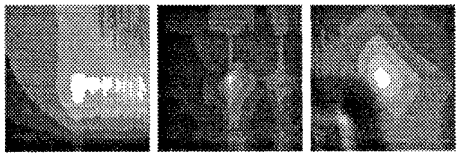

Figure 2: The autocorrelation surfaces of models European no-entry, Stop and Priority with contraction parameters s_x and s_y increasing from the lower left to the top right of the surfaces. The white centers of the autocorrelation surfaces are the correlation peaks and correspond to the contractional coherence areas of the signs. The European no-entry sign's contractional coherence area is much greater than the Stop sign's, which means that the contractional parameters s_x and s_y are easier to resolve for the Stop sign.

all information pertinent to the object's recognizability that is not contained in its edges. The danger in edge-based methods, therefore, is that a potentially larger amount of information may come from slowly varying brightness changes accumulated throughout the object's area than from rapid changes at edges. In this case, edge-based recognition methods are inherently sub-optimal. Conversely, if the predominant positional information about an object is concentrated in its edges, the foregoing analysis of Fisher information, coherence scales and complexity remains equally pertinent regardless the method of recognition. Moreover, the foregoing analysis goes beyond consideration of positional variations, as expressed in terms of the horizontal and vertical gradient components also used in edge-methods, but also accounts for the general linear variations permissible in an affine transformation.

7 Maximum likelihood estimation of an object in a scene image

In this section, we derive a method for recognizing an object imaged in a complex scene that attains the theoretical lower bound on mean-square error for any unbiased estimate, and therefore is by definition statistically optimal and information-conserving.

Given image data $\mathbf{I}$, and following classical estimation theory, we use the likelihood function of Eq. 3, to derive the maximum likelihood estimate $\hat{\mathbf{a}}_{ML} = \arg\max_{\mathbf{a}} P(\mathbf{I}|\mathbf{a})$ of the parameters $\mathbf{a}$, which can be found by solving the likelihood equation $\frac{\partial \ln P(\mathbf{I}|\mathbf{a})}{\partial \mathbf{a}}|_{\mathbf{a}=\hat{\mathbf{a}}_{ML}} = 0$. Since $\frac{\partial \ln P(\mathbf{I}|\mathbf{a})}{\partial \mathbf{a}} = \frac{\partial}{\partial \mathbf{a}}\left(-\frac{1}{2\sigma^2}\sum_k^{MN}(I_k - m_k(\mathbf{a}))\right)^2$, the maximum likelihood estimate

$$\hat{\mathbf{a}}_{ML} = \arg\min_{\mathbf{a}} \tfrac{1}{2\sigma^2}\Big(\sum_{(x,y)\in B}(I(x,y) - m(x,y))^2 + \sum_{(x,y)\in O^+}(I(x,y) - q(x,y;\mathbf{a}))^2\Big), \quad (23)$$

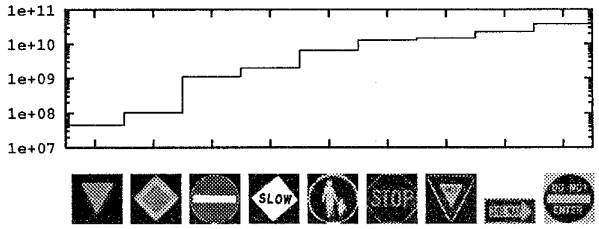

Figure 3: Comparison of complexity C for various traffic signs: Signs with inscriptions and human figures have higher complexity than signs composed only of simple geometric shapes. Our data analysis shows that the ability to unambiguously resolve a sign increases with the sign's complexity.

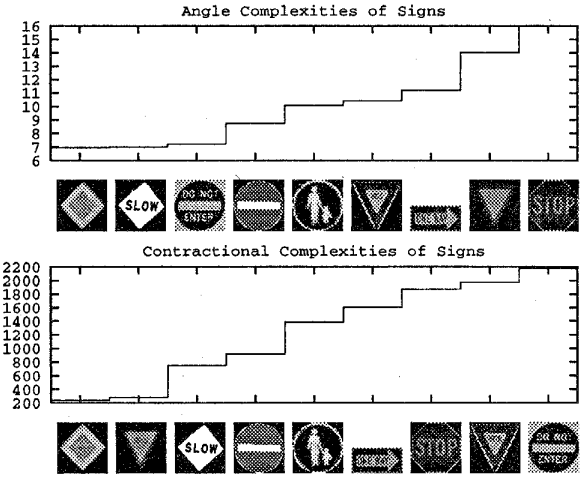

Figure 4: The rotational and contractional complexities of the traffic sign models.

where region B consists of background unrelated to the object while region O^+ is the union of all pixels that contain the expected object $q(x,y;\mathbf{a})$ as well as a slightly perturbed or variational object $q(x,y;\mathbf{a}+\Delta\mathbf{a})$. The first sum in Eq. 23 can be discarded, because the background does not depend on the object properties described by $\mathbf{a}$. The maximum likelihood estimate is then $\hat{\mathbf{a}}_{ML} = \arg\min_{\mathbf{a}} \sum_{(x,y)\in O^+}(I(x,y) - q(x,y;\mathbf{a}))^2$. After expanding the square, this reduces to

$$\hat{\mathbf{a}}_{ML} \approx \arg\max_{\mathbf{a}} \sum_{(x,y)\in O^+} I(x,y)q(x,y;\mathbf{a}), \quad (24)$$

because the data energy $\sum_{(x,y)\in O^+}(I(x,y))^2$ is always independent of $\mathbf{a}$, and, for small perturbations of $\mathbf{a}$ about its true value, the expected object energy $\sum_{(x,y)\in O^+}(m(x,y;\mathbf{a}))^2$ can be taken as a constant independent of $\mathbf{a}$.

We interpret $q(x,y;\mathbf{a})$, in Eq. 24, as a multidimensional matched filter, which, when evaluated at some

particular $\mathbf{a}_0$, is referred to as the *replica* $q(x,y;\mathbf{a}_0)$. To find the maximum likelihood estimate $\hat{\mathbf{a}}_{ML}$, therefore, we search for the replica that best matches with the object in the scene image. The value of the parameter vector that corresponds to this best match is the maximum likelihood estimate. It maximizes the output of the multidimensional matched filter given by the sum in Eq. 24.

To ensure that our filter is not biased by changes in either the data energy or the expected object energy within the local replica window, we employ a local weighting. The output of the resulting weighted multidimensional matched filter $r(\mathbf{a}) =$

$$\frac{A(\mathbf{a})}{\sigma_I(\mathbf{a})\sigma_q(\mathbf{a})}\left(\sum_{(x,y)\in O} I_q(x,y)q(x,y;\mathbf{a}) - m_I(\mathbf{a})m_q(\mathbf{a})\right), \tag{25}$$

quantifies how well the measured data in subimage $I_q(x,y)$ matches the replica object in $q(x,y;\mathbf{a})$. Here $A(\mathbf{a})$ is the number of pixels in the replica image $q(x,y;\mathbf{a})$ that have nonzero brightness, and therefore constitute the replica object, while O is the region that contains the replica object. The local variance of subimage $I_q(x,y)$ is $\sigma_I^2(\mathbf{a}) = A(\mathbf{a})\sum_{(x,y)\in O(\mathbf{a})} I_q(x,y)^2 - \left(\sum_{(x,y)\in O} I_q(x,y)\right)^2$, the variance of replica image $q(x,y;\mathbf{a})$ is $\sigma_q^2(\mathbf{a}) = A(\mathbf{a})\sum_{(x,y)\in O} q(x,y;\mathbf{a})^2 - \left(\sum_{(x,y)\in O} q(x,y;\mathbf{a})\right)^2$, and the local means are $m_I(\mathbf{a}) = \sum_{(x,y)\in O} I_q(x,y)$ and $m_q(\mathbf{a}) = \sum_{(x,y)\in O} q(x,y;\mathbf{a})$. It is noteworthy that the weighted multidimensional matched filter is dimensionless, with $|r(\mathbf{a})| \leq 1$ by the Cauchy-Schwartz inequality, so that scene object I_q and replica object q are perfectly correlated when $r(\mathbf{a}) = 1$.

When the estimate $\hat{\mathbf{a}}$ is very near to its true value, small changes in $\mathbf{a}$ lead to negligible changes in m_I, m_q, σ_I, and σ_q, so that these sample means and standard deviations may be taken as locally constant. In this case, the weighted matched filter of Eq. 25 becomes a linear function of the matched filter, the sum in Eq. 24, so that the value of $\mathbf{a}$ that maximizes $r(\mathbf{a})$ is the maximum likelihood estimate, where

$$\hat{\mathbf{a}}_{ML} = \arg\max_{\mathbf{a}} r(\mathbf{a}). \tag{26}$$

Therefore, while there may be local optima in the weighted matched filter output, the location of its global optimum in the parameter search space corresponds to the maximum likelihood estimate. The maximum likelihood estimate, and hence the weighted matched filter, asymptotically attains the lower bound derived in Eq. 10, and is therefore optimal and information-preserving when the *signal-to-noise-ratio* (SNR) E/σ^2 is high, as it is for the recognition problem with CCD data [2]. An experimental proof of our method's statistical optimality is readily provided by inspection of Fig. 5. Over the entire global peak, the weighted matched filter, a comparison of noiseless object replicas versus noisy image data, is indistinguishable from the autocorrelation of the coincident noiseless object replicas. This confirms that our approach is information-conserving, and therefore takes optimal advantage of all the image data pertinent to the object's recognizability.

8 Brightness invariance of flat surfaces

The brightness of an object depends on its reflectance properties, its shape, and its illumination. In particular, the *scene radiance* L of a surface patch centered at world point (X,Y,Z) is proportional to the *image irradiance* or *intensity* W measured at the corresponding pixel (x,y), such that $W(x,y) = gL(X,Y,Z)$, where g is a function of parameters of the imaging system [5]. Since the sensitivity of our imaging system is uniform over the whole image, we can assume that g is constant. The scene radiance is related to the object's bidirectional reflectance distribution function f_r and the source irradiance E_i by $L_r(X,Y,Z) = f_r(\mathbf{s}(X,Y,Z),\mathbf{v}(X,Y,Z),X,Y,Z)E_i(\mathbf{s}(X,Y,Z))$, where $\mathbf{s}(X,Y,Z)$ is the direction of a collimated light source, and $\mathbf{v}(X,Y,Z)$ is the direction of the camera. For a flat surface, however, the direction of the collimated source is constant over the object such that $\mathbf{s} = \mathbf{s}(X,Y,Z)$. Under the benign assumption that the object's reflectance has directional properties that are separable from its spatial properties, we have $f_r(\mathbf{s},\mathbf{v}(X,Y,Z),X,Y,Z) = f_{r1}(\mathbf{s},\mathbf{v}(X,Y,Z))\ \varrho(X,Y,Z)$. A special case of this is a Lambertian surface where $f_{r1}(\mathbf{s},\mathbf{v}) = 1/\pi$ and $\varrho(X,Y,Z)$ is the albedo. If the camera is at least a few object lengths away then its directional variations over the object will be so small that the camera's direction can be considered constant such that $\mathbf{v} = \mathbf{v}(X,Y,Z)$. Then the image brightness $I = W^{-\gamma}$ becomes

$$I = c_r\,\varrho^{-\gamma}(X,Y,Z), \tag{27}$$

which, to within the constant factor $c_r = (gf_{r1}(\mathbf{s},\mathbf{v})E_i(\mathbf{s}))^{-\gamma}$, is invariant to changes in the geometry of the source, receiver and object. It is noteworthy that in the case of a Lambertian surface, the above result is valid regardless of whether $\mathbf{v}(X,Y,Z)$ is effectively constant or not. By distributivity, these results are easily extended to a hemispherical distribution of distant sources, such as the sky, so that the

image brightness of the flat object remains invariant to changes in the geometry of the source, receiver and object to within the constant factor c_r.

9 Recognition of flat objects

The output of the weighted matched filter, given in Eq. 25, is invariant to linear transformations of image brightness of the form $I'(x,y) = c_1 I(x,y) + c_2$, where c_1 and c_2 are scalar constants [2]. But the analysis of the previous section shows that, to within a scalar factor, the image brightness of a flat object remains invariant to changes in scene shading brought upon by changes in the geometry of the source, receiver and object. The output of our weighted matched filter, therefore, is invariant to such changes in scene shading, as is our optimal estimate of the parameters **a** and necessary for object recognition.

10 Experimental results

Our data consists of more than 3280 complex real-world images. The details of the implementation of the system and related references can be found in Refs. [1, 2]. The system recognizes 94% of the traffic signs correctly and misclassifies 6%. The system performance depends on the complexity of the signs in the scene images. For example, the low-complexity European no-entry and European yield signs generally result in high filter outputs for arbitrary scenes and are therefore responsible for most of the false matches. Conversely, traffic signs with inscriptions and complicated shapes are generally more complex and easier to unambiguously recognize. This fact can be used apriori in evaluating the cross-class performance of a recognition system.

11 Summary and conclusions

We have developed a general method for object recognition that is information-conserving, attains the theoretical lower bound on estimation error for any unbiased estimate regardless the method of estimation, and is therefore statistically optimal. Our work provides a foundation for quantitative comparisons between different recognition methods and shows under what special circumstances sub-optimal techniques, such as purely edge-based methods, can become optimal. We have applied our theoretical results to develop a system that has successfully recognized traffic signs in thousands of complex real-world scenes.

In future work, we will extend our approach to non-planar 3-D objects, using physical models [5, 9, 7] that describe the imaging process and the object's reflectance properties.

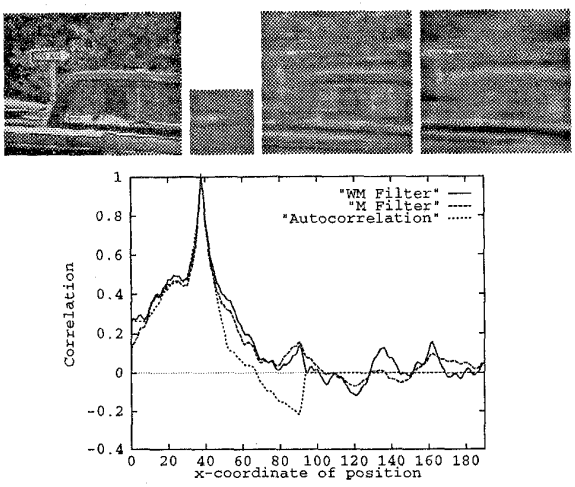

Figure 5: A scene image with a oneway sign and three ambiguity surfaces computed for all possible translations of the replica of a oneway sign with fixed angle and scaling parameters. The top right surface is the sign's autocorrelation. The middle left surface is computed using the matched (M) filter (Eq. 39), the middle right surface is computed using the weighted matched (WM) filter (Eq. 40). The correlation peak of the surfaces is a white spot located in the upper left of each plot. The graph shows horizontal slices through the global peaks of the ambiguity surfaces. The methods converge at the true solution.

References

[1] M. Betke and N. C. Makris. Fast object recognition in noisy images using simulated annealing. In *Proc. International Conference on Computer Vision*, pages 523–530, 1995. (MIT AI Memo 1510.).

[2] M. Betke and N. C. Makris. Information-conserving object recognition. Technical Report CS-TR-3799, 1997. http://www.cs.bc.edu/~betke.

[3] J. V. Difranco and W. L. Rubin. *Radar Detection.* Prentice-Hall, 1968.

[4] D. Gabor. Theory of communication. *J. Inst. Electri. Eng.*, 93:429–457, 1946.

[5] B. K. P. Horn. *Robot Vision.* MIT Press, 1986.

[6] N. C. Makris. A foundation for logarithmic measures of fluctuating intensity in pattern recognition. *Optics Letters*, 20:2012–2014, 1995.

[7] S. K. Nayar and M. Oren. Visual appearance of matte surfaces. *Science*, 267:1153–1156, 1995.

[8] G. Strang. *Linear Algebra and its Applications.* Academic Press, 1976.

[9] K. Torrance and E. Sparrow. Theory for off-specular reflection from rough surfaces. *J. Opt. Soc. Am. A*, 1:1105–1114, 1967.