International Journal of Computer Vision 44(1), 5–40, 2001

Recognition, Resolution, and Complexity of Objects Subject to Affine Transformations*

MARGRIT BETKE
Boston University, Boston, MA 02215
betke@cs.bu.edu, www.cs.bu.edu/faculty/betke

NICHOLAS C. MAKRIS
Massachusetts Institute of Technology, Cambridge, MA 02139

Received July 16, 1997; Revised March 8, 2001; Accepted March 8, 2001

Abstract. The problem of recognizing objects subject to affine transformation in images is examined from a physical perspective using the theory of statistical estimation. Focusing first on objects that occlude zero-mean scenes with additive noise, we derive the Cramer-Rao lower bound on the mean-square error in an estimate of the six-dimensional parameter vector that describes an object subject to affine transformation and so generalize the bound on one-dimensional position error previously obtained in radar and sonar pattern recognition. We then derive two useful descriptors from the object's Fisher information that are independent of noise level. The first is a generalized coherence scale that has great practical value because it corresponds to the width of the object's autocorrelation peak under affine transformation and so provides a physical measure of the extent to which an object can be resolved under affine parameterization. The second is a scalar measure of an object's *complexity* that is invariant under affine transformation and can be used to quantitatively describe the ambiguity level of a general 6-dimensional affine recognition problem. This measure of complexity has a strong inverse relationship to the level of recognition ambiguity. We then develop a method for recognizing objects subject to affine transformation imaged in thousands of complex real-world scenes. Our method exploits the resolution gain made available by the brightness contrast between the object perimeter and the scene it partially occludes. The level of recognition ambiguity is shown to decrease exponentially with increasing object and scene complexity. Ambiguity is then avoided by conditioning the permissible range of template complexity above a priori thresholds. Our method is statistically optimal for recognizing objects that occlude scenes with zero-mean background.

Keywords: object recognition, object complexity, coherence scale, coherence volume, recognition ambiguity, Fisher information, bandwidth, statistical estimation theory, lower bounds on estimation error, background models, traffic sign recognition, simulated annealing

1. Introduction

The problem of recognizing objects in complex real-world scenes is investigated for objects that can be uniquely determined by the six parameters of an affine transformation as well as a seventh parameter that identifies the object class. Experimental data is used to determine the joint probability distribution of the pixel brightness measurements in our charge-coupled device (CCD) images, which we find are corrupted by zero-mean, additive Gaussian noise. This distribution is then

*The support of the Office of Naval Research and the National Science Foundation is gratefully acknowledged.

used to construct the likelihood function for the affine parameter vector that defines the object to be estimated from our image data.

Two useful descriptors of an imaged object that are independent of noise level are derived from the object's Fisher information, which can be computed directly from the likelihood function. The first is a generalized coherence scale that determines the extent to which an object is self-correlated under affine transformation and so provides a physical measure of the extent to which an object can be resolved under affine parameterization. The second is a scalar measure of an object's *complexity* that is invariant under affine transformation and has a strong inverse relationship to recognition ambiguity. The practical value of this measure of complexity is that the ambiguity level of the recognition problem can be quantitatively described by it. We also derive the general Cramer-Rao lower bound on the mean square error in an estimate of the six-dimensional affine parameter vector that describes the 2-D position, rotation, dilation, and skew of an object in a zero-mean scene with additive noise and so generalize the bound on one-dimensional position error derived previously in radar and sonar pattern recognition (Van Trees, 1968; Levanon, 1988).

To address the problem of recognizing objects in complex real-world scenes that generally contain nonzero-mean backgrounds we develop a recognition method based on the normalized correlation coefficient. The coefficient is used as a "match measure" (Rosenfeld and Kak, 1982) between some portion of the scene and a "template object." The template object is computed by an affine transformation of its corresponding "model image." Model images are collected in advance and represent the classes of objects to be recognized. Our method searches for the affine parameter set that describes the two-dimensional rigid body motion and linear distortion that the model object must undergo in order to correlate with the scene object. The object is considered recognized if the normalized correlation coefficient reaches a predefined threshold for this parameter set. The relationship between the normalized correlation coefficient, the matched filter, and the maximum likelihood estimator is discussed.

Since the recognition problem is inherently nonlinear, a global optimization procedure is necessary for its solution. We develop a global search method based on simulated annealing. Our method's performance is evaluated experimentally by applying it to the problem of recognizing traffic signs in images of complicated outdoor scenes. For flat objects, such as traffic signs, we show that our affine parameterization is sufficient for recognition, so long as the objects do not have purely specular surfaces. For inherently three-dimensional objects, the parameter vector must be supplemented to 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 traffic sign case, however, we show that the normalized correlation coefficient is invariant to the uniform variations in shading characteristic of the signs.

In both our theoretical and experimental analysis, we find that the level of ambiguity, as measured by the number of incorrect matches, is strongly dependent upon both our measures of the complexity of the object and the complexity of the background scene. In general, we find that the level of ambiguity falls off with an exponential trend as complexity increases. To prevent false matches, we then find that it is necessary to precondition the recognition system with sufficiently large background and template complexities. We also show that the ambiguity level cannot be meaningfully characterized solely by the relative size of a template or scene object.

2. Position Estimation in Nonzero-Mean Background

In this section, we use examples from 1-D position estimation to show that (1) nonzero-mean backgrounds affect object resolution and recognition ambiguity and (2) the normalized correlation coefficient rather than the matched filter is the appropriate tool for object recognition in nonzero-mean scenes.

References (Zadeh and Ragazzini, 1952; Difranco and Rubin, 1968; Van Trees, 1968; Rosenfeld and Kak, 1982) define the classical matched filter $h(x)$, for a 1D signal $s(x)$ that is corrupted by additive zero-mean noise $n(x)$, as the impulse response

$$h(x) = cs(x_m - x). \tag{1}$$

Here $h(x)$ is the signal shifted by its true position x_m and "time-reversed," and c is a constant invariant with respect to the true and test positions. The position estimate $\hat{x}_m$ is then given by the lag at which the matched

filter output

$$f(x) = c \int_{-L/2}^{L/2} h(x-\xi)\,(s(\xi)+n(\xi))\,d\xi$$
$$= c \int_{-L/2}^{L/2} s(\xi-(x-x_m))\,(s(\xi)+n(\xi))\,d\xi \quad (2)$$

is maximized, i.e., $\hat{x}_m = \arg\max f(x)$. The additive noise leads to fluctuations in the peak, so that $\hat{x}_m$ is not necessarily x_m. In the absence of noise, however, the matched filter reduces to an unnormalized autocorrelation of the signal with peak at x_m, since

$$f(x) = cR_s(x-x_m) + c \int_{-L/2}^{L/2} s(\xi-(x-x_m))\,n(\xi)\,d\xi \quad (3)$$

where $R_s(x)$ is the unnormalized autocorrelation of the signal $s(x)$. The peak output of the autocorrelation is at the true location x_m of the signal. It follows that in the case of high signal-to-noise ratio (SNR), the estimate $\hat{x}_m$ becomes unbiased. Moreover, the matched filter $f(x)$ is a sufficient statistic according to information theory, and for high SNR attains minimum variance, and is therefore asymptotically optimal according to classical estimation theory. Among all linear filters, its output also has maximum signal-to-noise ratio (Difranco and Rubin, 1968). The statistical optimality characteristics of the matched filter are discussed in more detail in Section 8. The classical approach is illustrated in Fig. 1, where an object occludes a zero-mean background in a scene with stationary additive noise. Since the given SNR is high, the position of the object can be estimated optimally with a matched filter.

For the object recognition problem in computer vision, the brightness value at each pixel in the background is often expected to be nonzero. This is illustrated in Fig. 2, which shows the same object as in Fig. 1 now occluding a nonzero-mean background. The classical matched filter is not designed for this case and gives the wrong position estimate as shown in Fig. 2(c). Even in the absence of noise, $n(x) = 0$, the matched filter reduces to a cross-correlation, between object $s(x)$ and object in the expected background $i(x)$, the maximum of which is not guaranteed to fall at the true location of the object x_m. The *normalized* cross-correlation, however, can be applied successfully to the computer vision problem of recognizing an object that occludes a nonzero-mean background, as we discuss in Section 8, where the local background data is used to compute the normalizing factor.

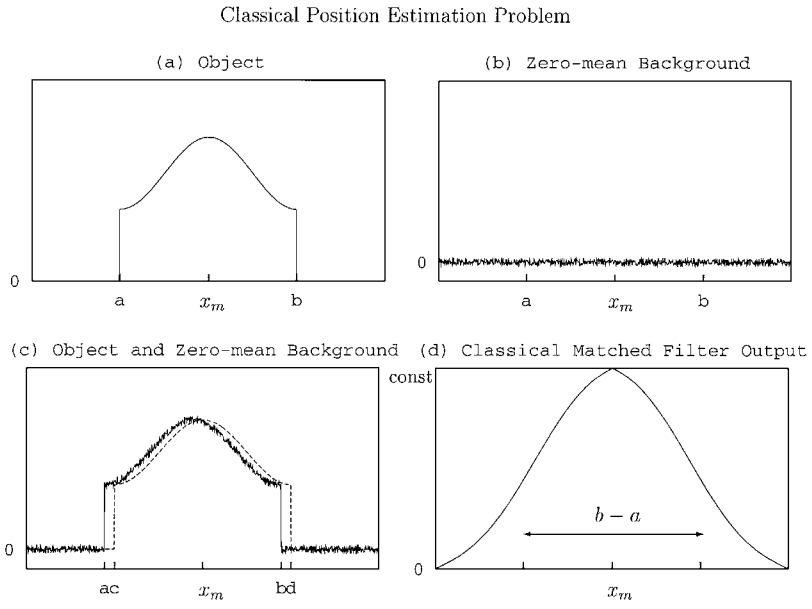


Figure 1. (a) The mean brightness values of an imaged object, located between a and b. (b) A scene with zero-mean background with stationary additive noise. (c) The object occludes the zero-mean background between a and b. A classical matched filter can be used to estimate the position of the object in this scene by shifting the object through the image and computing the filter output at each spatial lag. The dotted line in (b) shows the object at one of these lags, between c and d, where $d-c = b-a$. (d) The classical matched filter output. In high signal-to-noise, as in the case shown, the peak output is expected to occur at the true position x_m of the object.

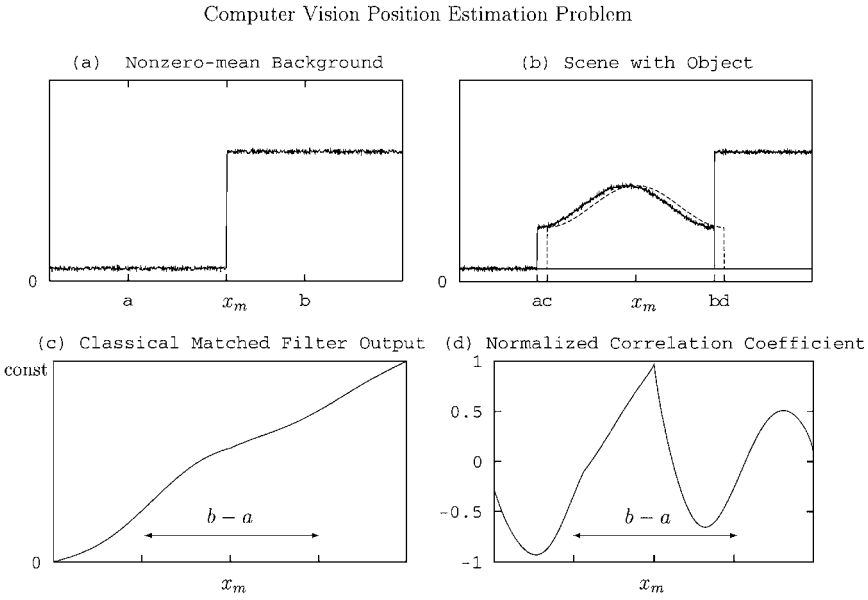


Figure 2. The graphs illustrate the computer vision scenarios. (a) A nonzero-mean background with zero-mean stationary additive noise $n(x)$. (b) Scene $i(x)+n(x)$, where the background in (a) is occluded by the object shown in Fig. 1(a). The dotted line shows a replica of the object that is shifted through the image and tested at each spatial lag to find the correct position x_m. (c) The output of the classical matched filter for the scene in (b). The peak output is not at the true position x_m, since the classical matched filter does not apply here. (d) The normalized correlation coefficient between the scene in (b) and the object shown in Fig. 1(a). The peak output is at the true position x_m.

Given a discontinuity between object and mean background, as in Fig. 2(b), the normalized cross-correlation not only yields a peak that converges to the true position of the object in high SNR, but also a position estimate of higher resolution than would be possible if the transition between object and background was continuous. This is illustrated in Fig. 2(d) where the filter's correlation peak is considerably sharper than the object's autocorrelation peak shown in Fig. 1(d). The gain in resolution follows from the local weighting which incurs a heavy penalty for the type of mismatch produced by a small shift away from the true position under such discontinuous transitions from object to background. Conversely, under smooth object-to-background transitions, or when the background along the perimeter of the object is constant, the normalized cross-correlation peak approaches that of the normalized autocorrelation, illustrated in Fig. 3(b), and a

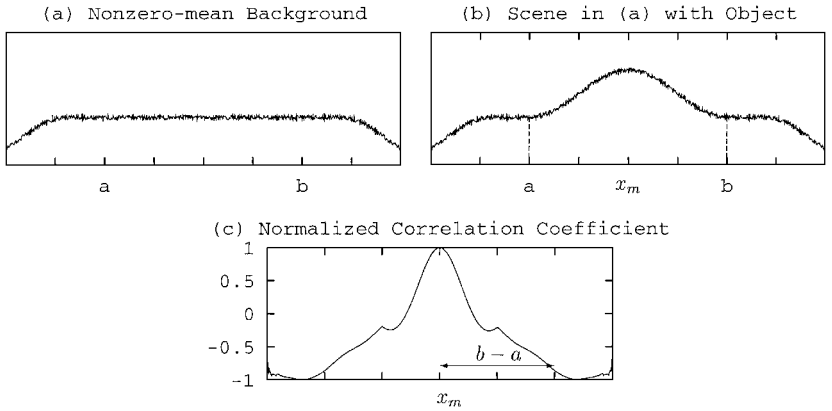


Figure 3. (a) A nonzero-mean background. (b) The same scene as (a), but part of the scene is occluded by the object shown in Fig. 1(a). (c) The normalized correlation coefficient between scene (b) and the object shown in Fig. 1(a) with a peak at x_m. Since the object-to-background transition is continuous, lower-resolution estimation than in the case shown in Fig. 2(b) results.

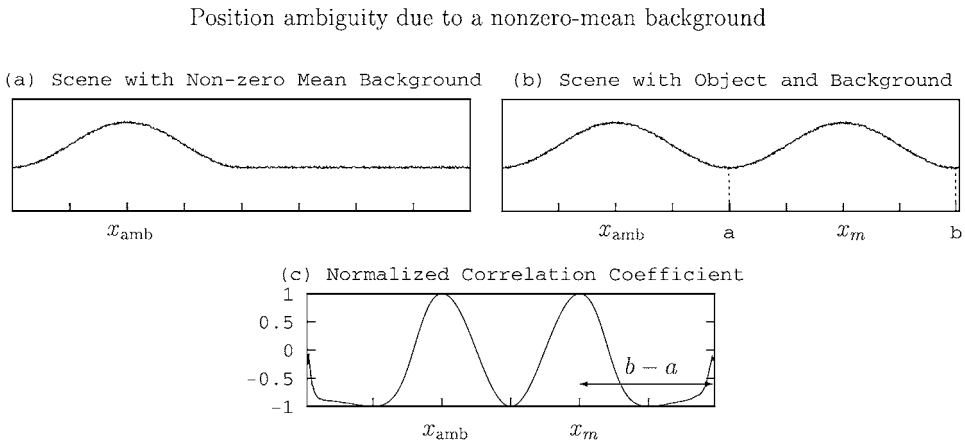


Figure 4. (a) A nonzero-mean scene that, on the left, has background ambiguity with shape identical to the object in Fig. 1(a). (b) The same scene, but on the right, the object occludes the nonzero-mean background between a and b. (c) The normalized correlation coefficient between scene (b) and the object in Fig. 1(a). The position of the object in (b) cannot be estimated unambiguously, since the object could be found either on the left at x_{amb} or on the right at x_m. In typical computer vision scenarios, however, such ambiguity is unlikely, unless the complexity of the scene object is low.

lower-resolution estimation of the object's position results. The crucial issue here is that the expected background can affect the form and resolution of an estimator. This effect does not appear in classical matched filter theory where the expected background is assumed to be zero.

The issue of deterministic background *ambiguity* also distinguishes the computer vision problem from the radar/sonar problem where the classical matched filter applies. In the latter, noise may cause spurious matches, but these become increasingly weak as the signal-to-noise ratio increases. In the former, the background scene is not zero mean, and false matches may occur even in the limiting case of no noise. This is illustrated in Fig. 4, where the background contains a feature with shape identical to that of the object given in Fig. 1(a). When the object is placed in this scene, as in Fig. 4(b), the optimal estimator localizes two objects, one at the correct position $x_{\dot{m}}$ and one at the position of the background ambiguity x_{amb}. Classical estimation theory offers no solution to this dilemma since it is based solely upon *statistical* optimality criteria, while the ambiguity here is deterministic. Moreover, these statistical criteria only apply when the estimate is in close proximity to the true solution and so do not touch upon the ambiguity issue.

Ambiguity then only becomes a serious problem when the object to be recognized is highly correlated with a portion of the background. In later sections, we define a quantitative measure of complexity, and show that such false matches are only likely to occur in the recognition of objects of low complexity. We also show that ambiguities can be avoided by preconditioning the recognition system's range of permissible template and background complexities.

3. The Statistics of Image Brightness

Charge-coupled device (CCD) cameras do not output the intensity W of light. Instead, they output a power-transformed intensity on an 8-bit grey-scale which we refer to as image *brightness* $I(x, y)$. The brightness is linearly proportional to $W^{-\gamma}(x, y)$ where γ is a "gamma correction," e.g., $\gamma = 2.2$ (Poynton, 1993). The purpose of this transformation is to correct for the response of cathode-ray tube monitors so that the output of any monitor is proportional to intensity.

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, as shown in Fig. 5, does not depend on the mean or on position (x, y). 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 Appendix A.

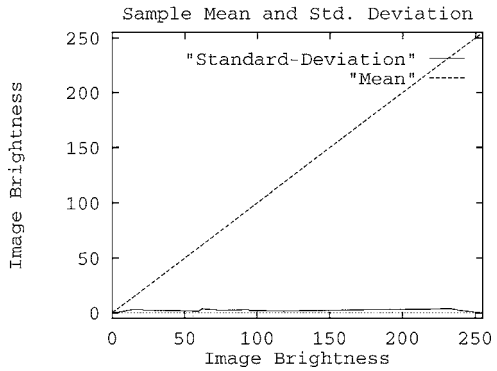


Figure 5. The measured mean and standard deviation of the image brightness I as a function of the mean. The sample standard deviation is signal independent and obtained by averaging hundreds of images of outdoor scenes. The average standard deviation is 2.65.

Our measured average skew of -0.02 and kurtosis of 2.81 are so close 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 concatenated 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 approximately

$$P(\mathbf{I}) = \frac{1}{(2\pi\sigma^2)^{MN/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^{MN} (I_k - m_k)^2 \right) \tag{4}$$

for $0 \leq I_k \leq \infty$, where the variance is constant and the mean varies throughout the image. If signal-dependent noise had been found, a nonlinear transformation of I could have been used to obtain signal-independent noise (Makris, 1995). Many references (Rosenfeld and Kak, 1982; Jain et al., 1995; Horn, 1886; Trucco and Verii, 1998; Umbaugh, 1998) also propose a Gaussian distribution to describe common noise in images and show how the noise can be removed to solve the classical image processing tasks of image restoration and enhancement. Our work, however, has a different focus. We model the noise so that we gain insight to the object recognition problem by considering it as a statistical parameter estimation problem.

4. 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 distortion 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 (x', y') are related to the rectangular Cartesian system (x, y) by the 2-D affine transformation

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \tag{5}$$

which can be expressed more succinctly as $\mathbf{x}' = \mathbf{A}\mathbf{x} - \mathbf{x_0}$, where

$$\mathbf{A} = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix} \begin{pmatrix} \cos\theta_0 & \sin\theta_0 \\ -\sin(\theta_0 + \alpha) & \cos(\theta_0 + \alpha) \end{pmatrix}. \tag{6}$$

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 from noisy image data.

5. 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 our case, the likelihood

function for $\mathbf{a}$, given the image data $\mathbf{I}$, is

$$P(\mathbf{I} \mid \mathbf{a}) = \frac{1}{(2\pi\sigma^2)^{MN/2}} \times \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{MN}(I_k - m_k(\mathbf{a}))^2\right) \quad (7)$$

where the mean $m_k(\mathbf{a})$ explicitly depends on the parameters to be estimated, while the noise variance σ^2 is independent of the parameter set. The form of the likelihood function in this physical approach is not arbitrary, but depends upon the probability distribution of the brightness measurements obtained with a CCD camera, and is very different from that found for other image data. For example, in many systems that employ phased arrays, such as microwave radar and active sonar, including medical ultrasound, or that employ coherent optical sources such as lasers, nonlinear speckle noise corrupts the resulting intensity images (Goodman, 1985; 1965; Makris 1996; Makris et al., 1995) due to fluctuations in the source, propagation medium, or scatterer. In these cases, the noise is not independent of the measured signal as it is for our image data. A homomorphic transformation of the measured data can then sometimes be used to transform the signal-dependent noise into signal-independent noise so that the object recognition techniques described in this paper can be applied (Makris, 1995; Downie and Walkup, 1994).

The Cramer-Rao lower bound (CRLB) on the mean-square error in any unbiased estimate $\hat{\mathbf{a}}$ can be expressed as

$$\mathrm{E}[(\hat{\mathbf{a}} - \mathbf{a})(\hat{\mathbf{a}} - \mathbf{a})^T] \geq \mathbf{J}^{-1}, \quad (8)$$

where the Fisher information matrix $\mathbf{J}$ is defined by

$$J_{ij} = -\mathrm{E}\left[\frac{\partial^2}{\partial a_i \partial a_j} \ln P(\mathbf{I} \mid \mathbf{a})\right] = \frac{1}{\sigma^2}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\left(\frac{\partial m(x, y; \mathbf{a})}{\partial a_i}\frac{\partial m(x, y; \mathbf{a})}{\partial a_j}\right). \quad (9)$$

If we consider a zero-mean background scene, 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 then becomes

$$J_{ij} = \frac{1}{\sigma^2}\sum_{(x,y)\in O^+}\frac{\partial q(x, y; \mathbf{a})}{\partial a_i}\frac{\partial q(x, y; \mathbf{a})}{\partial a_j}. \quad (10)$$

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

$$\mathrm{E}[(\hat{a}_i - a_i)^2] \geq [\mathbf{J}^{-1}]_{ii} = \frac{\sigma^2}{E}\ell_i^2, \quad (11)$$

where the object *energy*

$$E = \sum_{(x,y)\in O}|q(x, y; \mathbf{a})|^2 \quad (12)$$

and the *coherence scale*

$$\ell_i = \left([\mathbf{J}^{-1}]_{ii}\frac{E}{\sigma^2}\right)^{\frac{1}{2}} \quad (13)$$

for each parameter a_i are physical descriptors of the object that are independent of the noise level. The coherence scale ℓ_i does not depend on the noise variance σ^2, because σ^2 is included in the expression for the Fisher information in Eq. (10), and dividing the Fisher information by σ^2 therefore factors out the noise in Eq. (13). To put our definition of ℓ_i into historical perspective, we note that it generalizes the coherence scale found for the 1-D position estimate of a deterministic signal in additive noise (Kay, 1993; Levanon, 1988) to the 6 coherence scales necessary to describe object recognition after affine transformation in additive noise. Our Eq. (11) reduces to the 1-D case described, for example, in Eq. (9.45) in Levanon (1988) and in Eq. (3.38) in Kay (1993).

The coherence scale ℓ_i in Eq. (13) 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 as a function of 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.

When all parameters are uncoupled and $\mathbf{J}$ is diagonal, the product of n_a coherence scales $\ell_1 \cdots \ell_{n_a}$ yields a *coherence volume* that is a scalar measure characterizing the combined n_a-dimensional variations of the object, where n_a is the length of $\mathbf{a}$. More generally, we define the coherence volume V in terms of the determinant $|\mathbf{J}|$ of the Fisher information matrix by

$$V = \left(\frac{E}{\sigma^2}\right)^{\frac{n_a}{2}}|\mathbf{J}|^{-\frac{1}{2}}. \quad (14)$$

The lower bound can then be written as

$$\mathbf{J}^{-1} = \mathbf{J}_{adj}\left(\frac{\sigma^2}{E}\right)^{n_a} V^2, \tag{15}$$

where $\mathbf{J}_{adj}$ is the adjugate matrix of $\mathbf{J}$ (Strang, 1976).

This definition of a generalized coherence scale has great practical value because it corresponds to the width of the object's autocorrelation peak under affine transformation, as we will demonstrate, and so provides a physical measure of the extent to which an object can be resolved under affine parameterization. The generalized coherence scale is a physical measure independent of noise level since all statistical quantities are factored out in our definition. It is significant that in the strictest sense, the coherence scale depends not only on the pixels values within the object, but also those directly bordering the object since the Fisher information is defined over O^+, as is consistent with the intuitive analysis presented in Section 2. Neglect of the external pixels, however, enables one to more generally characterize the inherent self-coherence of an object or class of objects. This is valuable in assessing resolution attainable in a recognition problem involving the object or class of objects before information about the background scene is available. When bordering pixels are included, the generalized coherence scale can only remain the same or decrease. For example, such a decrease in a 1-D object's coherence length scale would lead to the increase in positional resolution described in Section 2.

From the computer vision perspective, 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. Our approach and purpose therefore stands apart from Cernuschi-Frias et al.'s (1989). 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. (15) 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 appears only as a constant factor in an object's information volume or recognizability, regardless of the noise level. This is a consequence of the physical structure of the likelihood function.

5.1. Two-Dimensional Position Resolution

We first derive the lower bound on the error for any unbiased two-dimensional position estimate of an object with known rotation, contraction and skew. Note, even this extends the classical radar/sonar 1-D positional Cramer-Rao lower bound of Van Trees (1968). Given the true position $(a_1, a_2) = (x_0, y_0)$, the Fisher information matrix, with elements

$$J_{ij} = \frac{1}{\sigma^2}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} \times\left(\frac{\partial q(x-x_0,\ y-y_0)}{\partial a_i}\,\frac{\partial q(x-x_0,\ y-y_0)}{\partial a_j}\right), \tag{16}$$

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. (16) be replaced by a continuous double integral so that $q(x, y)$ and $Q(u, v)$ can be defined as Fourier transform pair

$$Q(u,\ v) = \int\int_{O^+} q(x,\ y)e^{-j2\pi(xu+yv)}dxdy \tag{17}$$

and

$$q(x,\ y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Q(u,\ v)e^{j2\pi(xu+yv)}dudv \tag{18}$$

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

$$B_x^2 = \frac{(2\pi)^2}{(\Delta x)^2 E}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2|Q(u,\ v)|^2 dudv, \tag{19}$$

a mean-square bandwidth B_y^2 in y,

$$B_y^2 = \frac{(2\pi)^2}{(\Delta x)^2 E}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v^2|Q(u,\ v)|^2 dudv, \tag{20}$$

and a cross-term

$$B_{xy}^2 = B_{yx}^2 = \frac{(2\pi)^2}{(\Delta x)^2 E}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uv|Q(u,\ v)|^2 dudv, \tag{21}$$

with the aid of Parseval's Theorem

$$(\Delta x)^2 E = \int\int_{O^+} |q(x,\ y)|^2 dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |Q(u,v)|^2 dudv. \quad (22)$$

These definitions for the object's mean-square spatial bandwidth can be considered as a 2-D generalization of those introduced for one-dimensional signal waveforms by Gabor (1946). 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 Makris (1995), while those in the signal processing literature are defined about some average frequency that approximates the carrier frequency for 1-D narrowband signals.

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

$$\mathbf{J}^{-1} = \frac{\sigma^2}{E}\mathbf{B}^{-1} = \frac{\sigma^2}{E}\begin{pmatrix} B_y^2 & -B_{xy}^2 \\ -B_{xy}^2 & B_x^2 \end{pmatrix} A_{x_0,\ y_0}^2, \quad (23)$$

where

$$A_{x_0,\ y_0} = |\mathbf{B}|^{-\frac{1}{2}} \quad (24)$$

is the *coherence area* of the object, which follows from Eq. (14), where $V = A_{x_0,\ y_0}$ for this 2-D scenario. For example, the lower bound for estimating x_0 is simply

$$\mathrm{E}[(\hat{x}_0 - x_0)^2] \geq J_{x_0}^{-1} = \frac{\sigma^2}{E}\ell_{x_0}^2, \quad (25)$$

where the coherence length scale in x is

$$\ell_{x_0} = \frac{B_y^2}{|\mathbf{B}|} = B_y^2 A_{x_0,\ y_0}^2, \quad (26)$$

and the lower bound for y_0 is

$$\mathrm{E}[(\hat{y}_0 - y_0)^2] \geq J_{y_0}^{-1} = \frac{\sigma^2}{E}\ell_{y_0}^2, \quad (27)$$

where the coherence length scale in y is

$$\ell_{y_0} = B_x^2 A_{x_0,\ y_0}^2. \quad (28)$$

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 (Difranco and Rubin, 1968). While the coherence length scales ℓ_{x_0} and ℓ_{y_0} could have been obtained directly from Eq. (13) without introducing the mean-square bandwidth concept, this would have circumvented both the historical perspective and an important physical interpretation.

The great value of this formulation is that it leads to measures of coherence that correspond well with what is measured in practice. This is illustrated in Figs. 6–8 where the derived coherence scales are shown to provide a good measure of the peak widths of the respective autocorrelations of a given object as a function of 1-D or 2-D position lag. From the pattern recognition perspective, these coherence scales are interpreted as inherent physical scales to which the position of an object can be well resolved.

More specifically, the coherence areas and length scales of two traffic signs, a stop sign and a European no-entry sign, are illustrated in Fig. 6. The figure shows the signs' 2D-autocorrelation surfaces with white centers that correspond to the sign's coherence areas. The coherence area of the stop sign, comprised of 26 pixels, is less than half the size of the 56-pixel coherence area of the European no-entry sign. In practice, this means that, of the two signs, the position of the stop sign can be better resolved, by more than a factor of two. Figure 6 also illustrates 1D-horizontal slices through the center (c_x, c_y) of the signs' autocorrelation surfaces, where y-positions are fixed, i.e., $y = c_y$, and x-positions vary. The 8-pixel coherence length l_x of the European no-entry sign and 4-pixels coherence length of the stop sign correspond as indicated to the widths of the autocorrelation peaks. Of the two signs, the stop sign's horizontal position can be resolved better, by roughly a factor of two, because of its narrower autocorrelation peak-width and shorter coherence length.

While the coherence area is invariant to changes in object rotation, the coherence lengths scales and bounds on position estimation error are not, as is shown in Appendix B by principal component analysis.

5.2. *Angular Resolution*

To investigate the angular resolution of an object, consider the case when only the rotation θ_0 of the object about some point in the image plane is unknown. By Eq. (13), the angular coherence scale for object rotation is

$$\ell_{\theta_0} = \left(\frac{E}{\sum_{(x,y)\in O^+}\left|\frac{\partial q(x,y)}{\partial \theta_0}\right|^2}\right)^{\frac{1}{2}}. \quad (29)$$

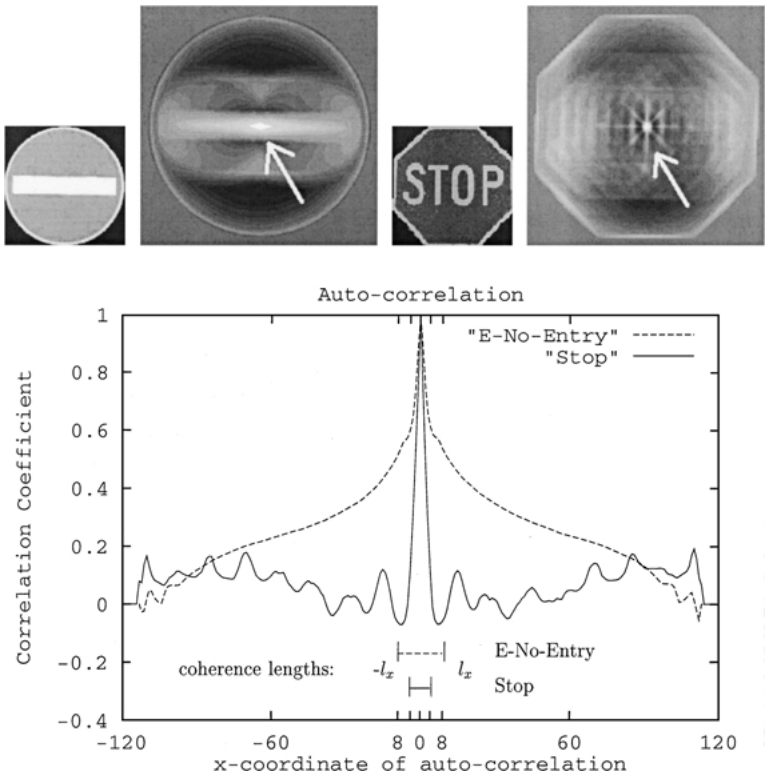


Figure 6. Above, the images of two traffic signs of size 120 × 120 and their 2D-autocorrelation surfaces are shown. The European no-entry sign has a coherence area of 56 pixels, the stop sign has a coherence area of 26 pixels. The white centers in the autocorrelation surfaces, indicated by arrows, correspond to the coherence areas of the signs. The position of the stop sign can then be resolved more easily than the position of the European no-entry sign. Below are 1D-horizontal slices through the center $(c_x,\ c_y)$ of the signs' autocorrelation surfaces, where y-positions are fixed, i.e., $y = c_y$, and x-positions vary. The coherence length l_x is 8 pixels for the European no-entry sign and 4 pixels for the stop sign. 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.

This leads to the bound

$$\mathrm{E}[(\hat{\theta}_0 - \theta_0)^2] \geq J_{\theta_0}^{-1} = \frac{\sigma^2}{E}\,\ell_{\theta_0}^2 \tag{30}$$

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 skew of the object, since E and ℓ_{θ_0} are functions of $s_x,\ s_y$, and α.

The present formulation again leads to a coherence scale that corresponds well with what is measured in practice. This is illustrated in Fig. 7 where the angular coherence scales of a stop sign, $\ell_{\theta_0} = 44^\circ$, and a European no-entry sign, $\ell_{\theta_0} = 20^\circ$, are shown to provide a good measure of the peak widths of the respective autocorrelation functions as a function of object rotation. The European no-entry sign has greater circular symmetry and therefore has a wider angular autocorrelation peak and a correspondingly larger angular coherence scale than the stop sign. The stop sign can then be better resolved in angle in a recognition problem.

5.3. *Contractional Resolution*

To investigate contractional resolution, consider the case when only an object's contractions 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. (10). The coherence area A_{s_x,s_y} and coherence length scales ℓ_{s_x}, ℓ_{s_y} are then dependent, by Eq. (14), on both diagonal and cross terms of the Fisher information matrix, such that

$$A_{s_x,s_y} = \left(\frac{E}{\sigma^2}\right)|\mathbf{J}|^{-\frac{1}{2}}, \tag{31}$$

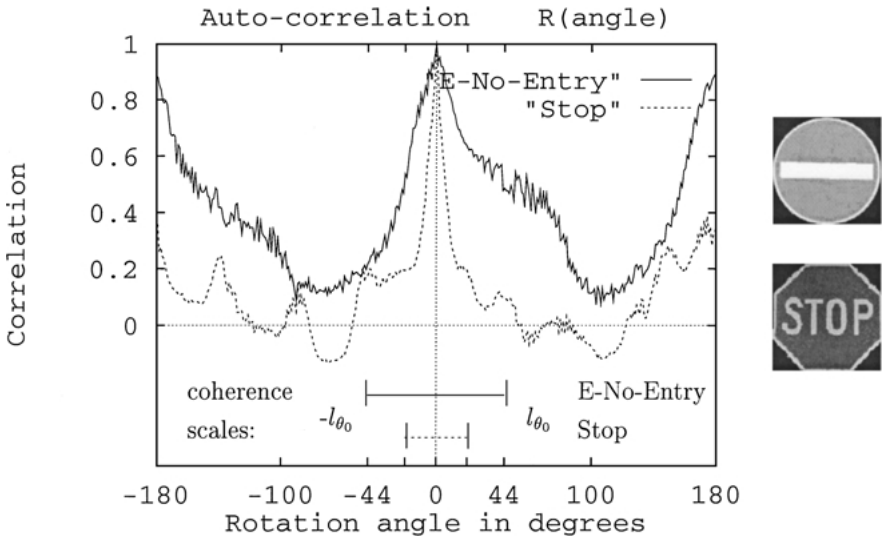


Figure 7. Comparison of angular coherence scales ℓ_{θ_0} and angular autocorrelations of two model signs. The European no-entry sign's angular coherence scale is $\ell_{\theta_0} = 44^\circ$. The stop sign's angular coherence scale is $\ell_{\theta_0} = 20^\circ$. The European no-entry sign's autocorrelation peak is much wider than the stop sign's, indicating that its rotation is more difficult to resolve.

$$\ell_{s_x} = \left([\mathbf{J}^{-1}]_{11} \frac{E}{\sigma^2} \right)^{\frac{1}{2}} \quad \text{and} \quad \ell_{s_y} = \left([\mathbf{J}^{-1}]_{22} \frac{E}{\sigma^2} \right)^{\frac{1}{2}}. \tag{32}$$

The bounds for contractional resolution are then

$$\mathrm{E}[(\hat{s}_x - s_x)^2] \geq \frac{\sigma^2}{E} \ell_{s_x}^2, \tag{33}$$

and

$$\mathrm{E}[(\hat{s}_y - s_y)^2] \geq \frac{\sigma^2}{E} \ell_{s_y}^2. \tag{34}$$

These scales and bounds are invariant to changes in object position but are only invariant to changes in object rotation when the contractions s_x and s_y are equal.

The present theory again leads to coherence scales that correspond well with what is measured in practice. This is illustrated in Fig. 8 where the contractional coherence areas and scales are shown to provide a good measure of the peak widths of the respective autocorrelations as a function of object contraction. Specifically, the figure shows results for a European no-entry sign, a stop sign, and a European priority sign. The contractional coherence areas of the signs are 0.009, 0.004, 0.029, respectively. Figure 8 also shows 1-D diagonal slices of the autocorrelation surfaces along the diagonal $s_x = s_y$. The coherence scales l_s of the European no-entry, stop, and European priority signs are 0.095, 0.06, and 0.17, respectively, indicating that the stop sign is most sensitive and can be resolved the best of the three under contraction, as might be expected from its complicated lettering.

6. 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 can help us 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 complexities 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. Because of 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 regard a coherence area as small if the ratio of it to the overall object area is much less than 1. Then, for example, the Mona Lisa's coherence area is small, due to its large number "of elaborately interconnected

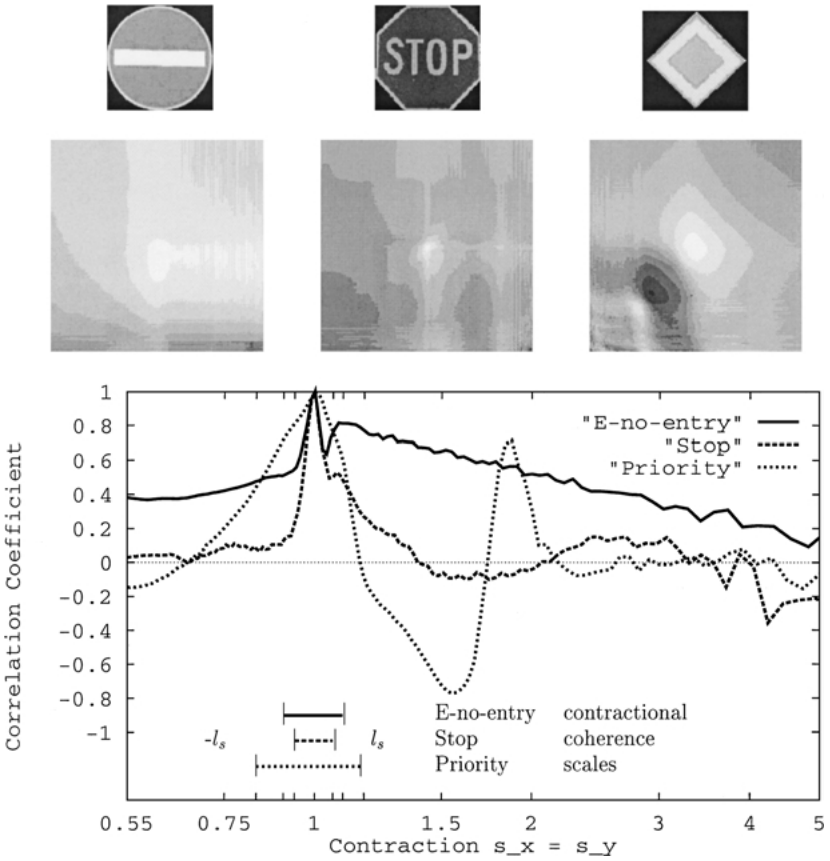


Figure 8. Above, the autocorrelation surfaces of model signs European no-entry, stop and priority are shown with contraction parameters s_x and s_y increasing from the lower left to the top right of the surfaces. The white centers of the autocorrelation surfaces are the correlation peaks and correspond to the contractional coherence areas of the signs, which are 0.009, 0.004, 0.029, respectively. Below, 1-D diagonal slices of the autocorrelation surfaces are shown along the diagonal $s_x = s_y$. The corresponding coherence scales l_s are 0.095, 0.06, and 0.17.

parts," but the number of coherence areas or coherence cells that fit into the Mona Lisa's overall area is large. Conversely, the coherence area of the blank canvas is not small, but the number of coherence cells that fit into the blank canvas' overall area is near unity. We may consider the overall object area as a kind of *outer scale* and the coherence area as a kind of *inner scale* for variations in an object's 2-D position. It is the ratio of this outer scale to its inner scale that determines the number of coherence cells or the degrees of freedom of the object. The higher the degrees of freedom, the more sensitive the object is to affine transformations and the easier it is to resolve or recognize. The degrees of freedom, so defined, serve 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 object area times $2\pi^2$. This is the product of the outer scales for 2-D positional transformation, rotation, 2-D contractions, and skew that respectively are 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. (14), so that

$$C = \frac{S}{V} = A\, 2\pi^2 \left(\frac{\sigma^2}{E}\right)^{\frac{n_a}{2}} |\mathbf{J}|^{\frac{1}{2}}. \qquad (35)$$

Equation (35) defines the complexity of an object using the determinant of its Fisher information matrix, which is given in Section 5, Eq. (10) for the case of zero-mean background scenes. The Fisher information matrix is defined for the set O^+ of pixels, which not only contains the pixels *comprising* the object, but also those zero-mean background pixels *bordering* the object. Including these neighboring pixels in the computation of

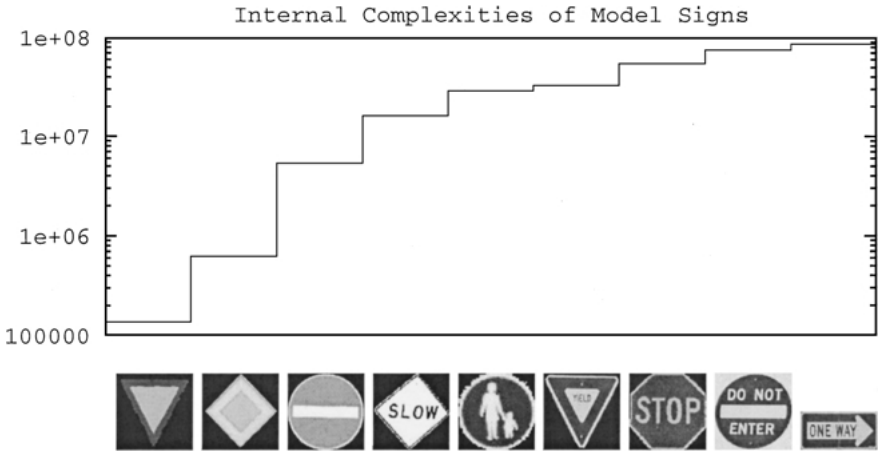


Figure 9. Comparison of the internal complexities C of various traffic signs: Signs with inscriptions and human figures have higher complexity than signs composed only of simple geometric shapes.

the Fisher information matrix is crucial, because their brightness values are affected by small changes in the parameter vector. An object that forms a strong contrast to the zero-mean background is easier to resolve than an object that forms a low contrast. Similarly, if we place an object in various nonzero-mean background scenes, its coherence volume and therefore complexity change from scene to scene, since the background bordering the object also changes from scene to scene. To provide a measure of an object's complexity that applies to the nonzero-mean case and does not vary with any particular scene background, it is practical to define the *internal complexity* using only the set O of pixels comprising the object and not its boundary. To distinguish it from the complexity defined in Eq. (35), we call the object's complexity based on the pixels in O^+ the *outer complexity*. The effect of external pixels on resolution and coherence discussed in Sections 2 and 5 is readily extended to the analysis of complexity, since complexity, by our definition, is simply inversely proportional to the generalized coherence scale.

To help fix ideas, consider again the illustrative examples presented in Section 2. The position of the object in Fig. 2(b) can be resolved better than the same object in Fig. 3(b) because of the discontinuous transition from object to background. This can also be interpreted directly in terms of the Fisher information which is defined in terms of partial derivatives of the brightness values of the scene with respect to the affine parameter vector describing the object. As a result, the coherence length of the object in the scene of Fig. 2(b) is smaller than in Fig. 3(b) if external pixels are included. While the inner complexities of the objects are the same in both scenes, the outer complexity of the object is then greater in Fig. 2(b).

The internal complexities of various traffic signs are compared in Fig. 9. As may be expected from a qualitative perspective, signs with inscriptions and human figures have much higher internal complexities than signs composed only of simple geometric shapes. The ability to unambiguously resolve an object in an arbitrary scene increases with the object's internal complexity, as is shown by data analysis in Section 15.1.

When the affine transformation is reduced to a 2-D translation, the relevant *positional complexity* becomes

$$C_{x_0,y_0} = \frac{A}{A_{x_0,y_0}}, \tag{36}$$

where the coherence area A_{x_0,y_0} is given in Eq. (24).

Similarly, we define the *rotational complexity* of an object by

$$C_{\theta_0} = \frac{2\pi}{\ell_{\theta_0}}, \tag{37}$$

and the *contractional complexity* by

$$C_s = \frac{1}{A_{s_x,s_y}}, \tag{38}$$

where the rotational coherence scale ℓ_{θ_0} is defined in Eq. (29) and the contractional coherence area A_{s_x,s_y} in Eq. (31). These positional, rotational, and contractional complexities of the traffic sign models are plotted in Fig. 10 and are consistent with qualitative appraisals

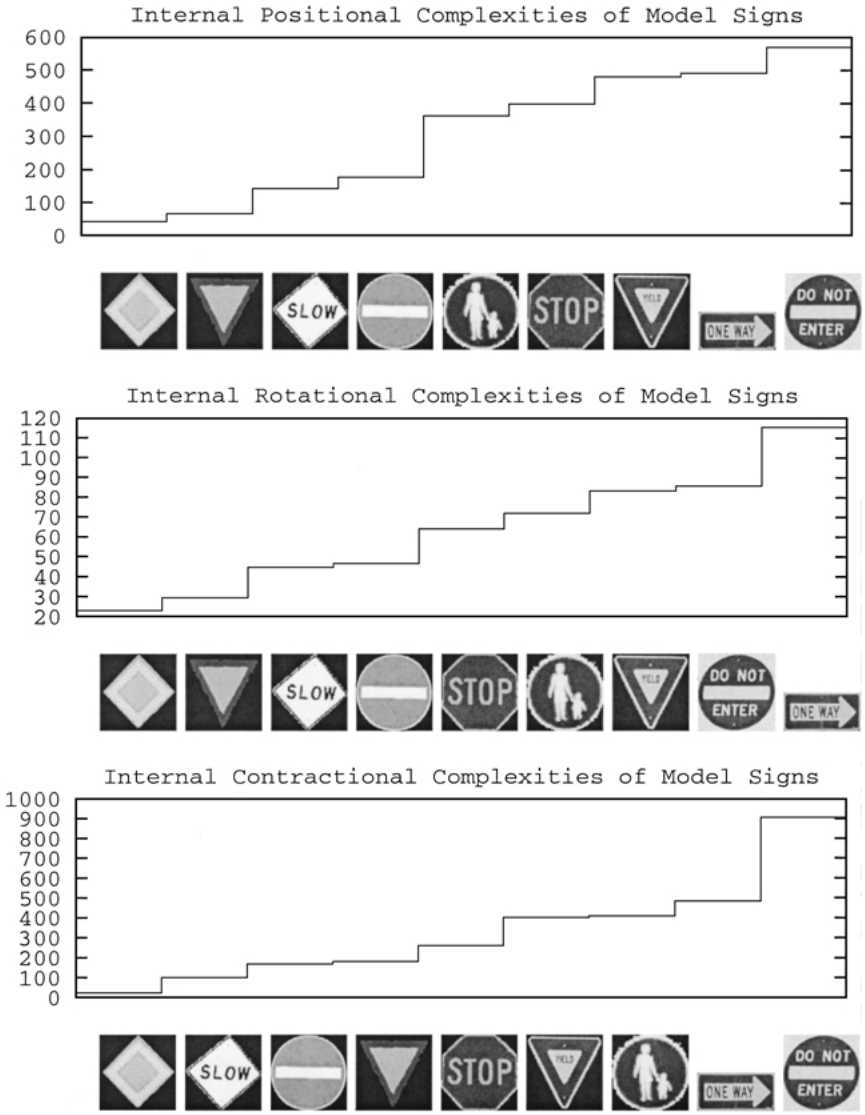


Figure 10. The positional, rotational, and contractional complexities of the traffic sign models.

of the inherent positional, rotational and contractional symmetries of the signs.

7. Image Edges

There is an important connection between the positional Fisher information of an object that occludes a zero-mean background 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. (16), 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 edges that comprise a relatively small fraction of the object's overall area. 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 all information pertinent to the object's recognizability that is not contained in its edges. For example, Fig. 11 shows the edge maps of an apricot for different thresholds. The danger in edge-based methods is that

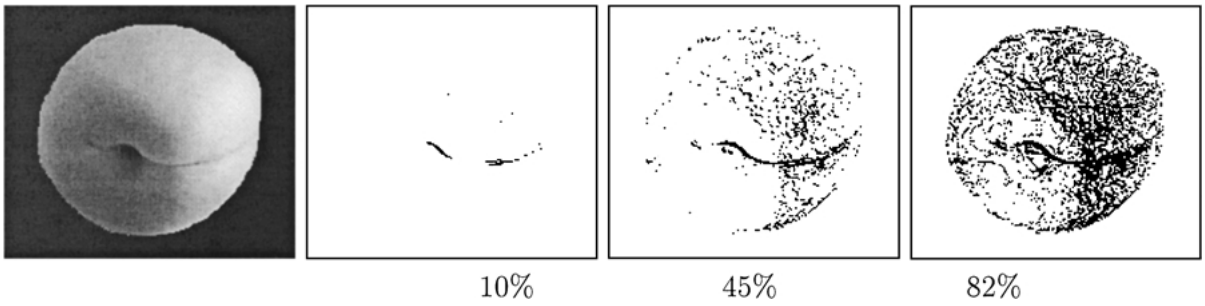


Figure 11. The image of an apricot and its edge maps computed for three different thresholds. Underneath the images, the positional complexities computed for the pixels that appear in these thresholded edge maps are given as an percentage of the complexity computed for all pixels.

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. Edge-based recognition methods are then inherently sub-optimal. This is the case for the apricot image in Fig. 11, which shows three different edge maps obtained from three different thresholds. The positional complexities computed for the respective edge maps are roughly a tenth, half, and 4/5 of the complexity for the original image. The higher the threshold becomes, the larger the loss of information and the lower the complexity of the resulting edge image become.

Conversely, if the predominant positional information about an object is concentrated in its edges, the analysis of Fisher information, coherence scales and complexity remains equally pertinent regardless of the method of recognition. Figure 12 shows the edge maps of a stop sign computed for the same thresholds as used in Fig. 11. The complexity analysis of the three edge maps shows that the positional information about the stop sign is contained in its edges, because the positional complexities computed for its respective thresholded edge maps is 90%, 95%, and 98% of the complexity for the original image.

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. Figure 13 illustrates such variations, showing not only the horizontal and vertical partial derivatives of the stop sign, but also its rotational and contractional derivatives.

Figure 12. A stop sign and three edge maps computed using same thresholds as in Fig. 11. The predominant positional information about the stop sign is concentrated in its edges, because the positional complexities computed for its respective edge maps is 90%, 95%, and 98% of the complexity of the original image.

8. Affine Parameter Estimation Using the Normalized Correlation Coefficient

In this section, we describe methods for parameter estimation in scenes corrupted by additive Gaussian noise and discuss the cases of objects occluding zero-mean and nonzero-mean backgrounds. We address the issues outlined in Section 2 and show how the methods are related and where they differ for both background scenarios.

Figure 13. A stop sign and its partial derivatives with respect to x, y, θ, and $s = s_x = s_y$.

8.1. Maximum Likelihood Estimation in Scenes With Zero-Mean Backgrounds

We first discuss parameter estimation in scenes corrupted by additive Gaussian noise where the object occludes a zero-mean background. If scene image I is described by the likelihood function in Eq. (7), the maximum likelihood estimate $\hat{\mathbf{a}}_{ML} = \arg\max_{\mathbf{a}} P(\mathbf{I} \mid \mathbf{a})$ can be derived by maximizing the argument $-\frac{1}{2\sigma^2}\sum_{(x,y)\in I}(I(x, y) - m(x, y; \mathbf{a}))^2$ of the exponent of the Gaussian distribution, since the noise variance σ^2 does not depend on parameter vector $\mathbf{a}$. This argument is also the Mahalanobis distance (Rao, 1973) between the image mean $m(x, y; \mathbf{a})$ that depends on $\mathbf{a}$ and the measured scene data $I(x, y)$. Solving $\frac{\partial}{\partial \mathbf{a}} \ln P(\mathbf{I} \mid \mathbf{a})|_{\mathbf{a}=\hat{\mathbf{a}}_{ML}} = 0$ yields the maximum likelihood estimate

$$\hat{\mathbf{a}}_{ML} = \arg\min_{\mathbf{a}} \sum_{(x,y)\in I} (I(x, y) - m(x, y;\ \mathbf{a}))^2. \quad (39)$$

The maximization is performed over the allowable range of values for $\mathbf{a}$ and can be implemented as an exhaustive search. The resulting maximum likelihood estimator (MLE) is unbiased and attains the Cramer-Rao lower bound for large datasets or high signal-to-noise ratios (SNR) according to classical estimation theory (Van Trees, 1968). For typical template objects obtained from low-variance CCD camera measurements, as the traffic signs in Fig. 9, the SNR is high so that the MLE is asymptotically optimal.

Equation (39) is also called the *sum-squared difference* (*SSD*) *measure of match* (Rosenfeld and Kak, 1982) and has been used extensively in computer vision to recognize objects in images with zero-mean backgrounds. In these applications, the task may be object *identification* as, for example, in printed character recognition. Then template and scene objects are assumed to have the same sizes, and object position and orientation are fixed. If the task is object *localization*, template object q is considered small in comparison to scene I and its position $\mathbf{a}$ in scene I is estimated. Here the pixels that comprise template $q(x, y; \mathbf{a})$ are the *nonzero* image mean pixels $m(\mathbf{a}, x, y)$.

By expanding the SSD to $\sum I(x, y)^2 - 2\sum I(x, y)\, q(x, y; \mathbf{a}) + \sum q(x, y; \mathbf{a})^2$, the measure of match $\sum I(x, y) q(x, y; \mathbf{a})$ (Rosenfeld and Kak, 1982; Ballard and Brown, 1982; Jain et al., 1995) can be derived as a measure of correlation between object and scene as a function of 2D position and rotation for fixed skew and scale. Since the energies $\sum I(x, y)^2$ and $\sum q(x, y; \mathbf{a})^2$ are constant for fixed skew and scale, maximizing the cross-correlation $\sum I(x, y) q(x, y; \mathbf{a})$ also maximizes the likelihood function P and therefore yields an asymptotically optimal estimate in this case. Template q can then be interpreted as a three-dimensional matched filter over 2-D position and rotation within the image plane when the object scale and skew are known.

8.2. Normalized Correlation Coefficient

To address the problem of recognizing objects in complex real-world scenes, we use a correlation measure that is not computed over the whole scene image as in the zero-mean background case, but instead compares the template object q only with a scene *subimage* I_q. Since various subimages must be tested during the search for the best matching template, the subimage I_q is not fixed, as in the zero-mean background case, where $I_q = I$. The square root of the energy $\sum I_q^2$ is then used as a normalizing factor so that matches with different subimages can be compared. The resulting *normalized cross-correlation* $(\sum I_q(x, y) q(\mathbf{a}, x, y)) / (\sum I_q(x, y)^2)^{\frac{1}{2}}$ serves as a match criterion for object identification and localization in many computer vision applications (e.g., Rosenfeld Kak, 1982; Kelley et al., 1983; Yoshimura and Kanade, 1994). If, in addition to object class and position, object orientation is also unknown, the normalized cross-correlation is computed for a set of templates, where each template has a different orientation (Kashioka et al., 1976; Chin and Dyer, 1986; Kelley et al., 1983; Yoshimura and Kanade, 1994; Rosenfeld and Kak, 1982).

An additional step is taken in Rosenfeld and Kak (1982) which proposes to normalize the cross-correlation by subtracting the average gray level in the image from the gray level of each pixel. The resulting match measure is the normalized correlation coefficient $r = (\sum (I_q - \hat{m}_{I_q})(q - \hat{m}_q)) / (\hat{\sigma}_{I_q}\, \hat{\sigma}_q)$, where $\hat{m}_{I_q}$ and $\hat{m}_q$ are the respective sample means of the subimage and template, and $\hat{\sigma}_{I_q}$ and $\hat{\sigma}_q$ the respective sample standard deviations. The normalized correlation coefficient is dimensionless, with $|r| \leq 1$ by the Cauchy-Schwartz inequality, so that scene object I_q and template object q are perfectly correlated when $r = 1$. Rosenfeld and Kak (1982) calls attention to the shift invariance of r, which means that two templates that differ from each other by a gray-scale shift yield the same normalized correlation coefficient r when matched with a scene.

The normalized correlation coefficient's shift invariance therefore provides an important advantage over other match measures. In Appendix C we generalize this argument and show the *linear* shift invariance of r, which allows recognition of a scene object whose brightness linearly differs from the brightness of the template object.

Most importantly, the normalized correlation coefficient can be used to estimate the full affine parameter vector $\mathbf{a}$ in nonzero-mean scenes. It therefore applies to the general object recognition problem of finding a translated, rotated, dilated, and skewed object in a scene by quantifying how well the measured data in subimage $I_q(x, y)$ matches the template object in $q(x, y; \mathbf{a})$. The local weighting by the standard deviations of scene and template images ensures that the coefficient is not biased by changes in either the scene energy or the expected object energy within the local template window. We choose a computationally advantageous form of the normalized correlation coefficient for our computer vision system:

$$r(\mathbf{a}) = \frac{1}{\hat{\sigma}_{I_q}(\mathbf{a})\hat{\sigma}_q(\mathbf{a})}\left(A(\mathbf{a})\sum_{(x,y)\in O} I_q(x, y)q(x, y; \mathbf{a}) - \hat{m}_{I_q}(\mathbf{a})\hat{m}_q(\mathbf{a})\right). \quad (40)$$

Here $A(\mathbf{a})$ is the number of pixels in the template image $q(x, y; \mathbf{a})$ that have nonzero brightness, and therefore constitute the template object, while O is the region that contains the template object, as illustrated in Fig. 14. The respective sample variances of subimage $I_q(x, y)$ and template image $q(x, y; \mathbf{a})$ are $\hat{\sigma}^2_{I_q}(\mathbf{a}) = A(\mathbf{a})\sum_{(x,y)\in O} I_q(x, y)^2 - (\sum_{(x,y)\in O} I_q(x, y))^2$ and $\hat{\sigma}^2_q(\mathbf{a}) = A(\mathbf{a})\sum_{(x,y)\in O} q(x, y; \mathbf{a})^2 - (\sum_{(x,y)\in O} q(x, y; \mathbf{a}))^2$, and the respective image sample means are $\hat{m}_{I_q}(\mathbf{a}) = \sum_{(x,y)\in O} I_q(x, y)$ and $\hat{m}_q(\mathbf{a}) = \sum_{(x,y)\in O} q(x, y; \mathbf{a})$.

The normalized correlation coefficient, SSD, and autocorrelation produce similar outputs over the parameter set, known as ambiguity surfaces, with strong peaks at the true value for $\mathbf{a}$ so long as object complexity is high and the background around the object perimeter is constant. Potential differences between the shapes of the global peaks are due to nonuniform object-background contrasts, which improve resolution, as shown in Section 2. Consider the ambiguity surface of Fig. 15 for the position estimate of an object that blends in relatively well with the background. For such cases, when the estimate $\hat{\mathbf{a}}$ is very close to its true value, small changes in $\mathbf{a}$ can lead to negligible changes in m_I, m_{I_q}, σ_I, and σ_{I_q}, which may then be taken as locally constant. Over the entire global peak, the normalized correlation coefficient, which compares a noiseless object replica to noisy image data, then becomes indistinguishable from the autocorrelation of the coincident noiseless object replicas.

The ambiguity surfaces shown in Fig. 15 are typical of those derived during the extensive experiments described in Section 15, where we find that high-complexity objects can be recognized to within a pixel width. This accuracy is consistent with the corresponding sub-pixel-width positional Cramer-Rao error bounds obtained for the high SNR encountered in the traffic sign recognition problem. This finding is significant because it means that the method of estimation has a variance that effectively attains the Cramer-Rao lower bound. According to classical estimation theory, any estimator that attains the Cramer-Rao lower

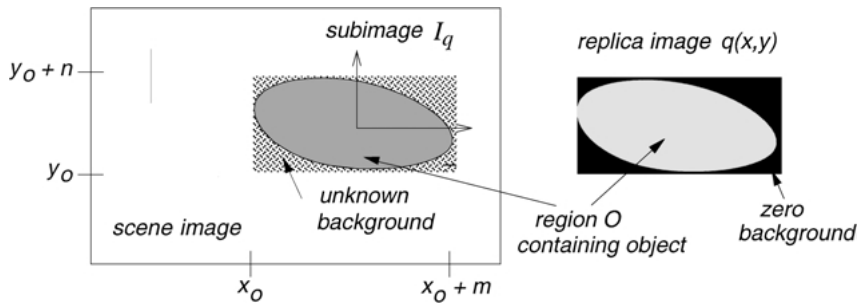


Figure 14. Scene image $I(x, y)$ with subimage $I_q(x, y)$ and replica or template image $q(x, y)$. Since the replica object may not be exactly rectangular, the portion $I_q(x, y)$ of the scene image that does not overlap the object replica must be removed from the match. To do so, the $m \times n$ replica image $q(x, y)$ is set to zero for pixels not belonging to the replica object. The computation time for any value of r is proportional to the number of nonzero pixels A in the object, which is usually much smaller than the number of pixels in I.

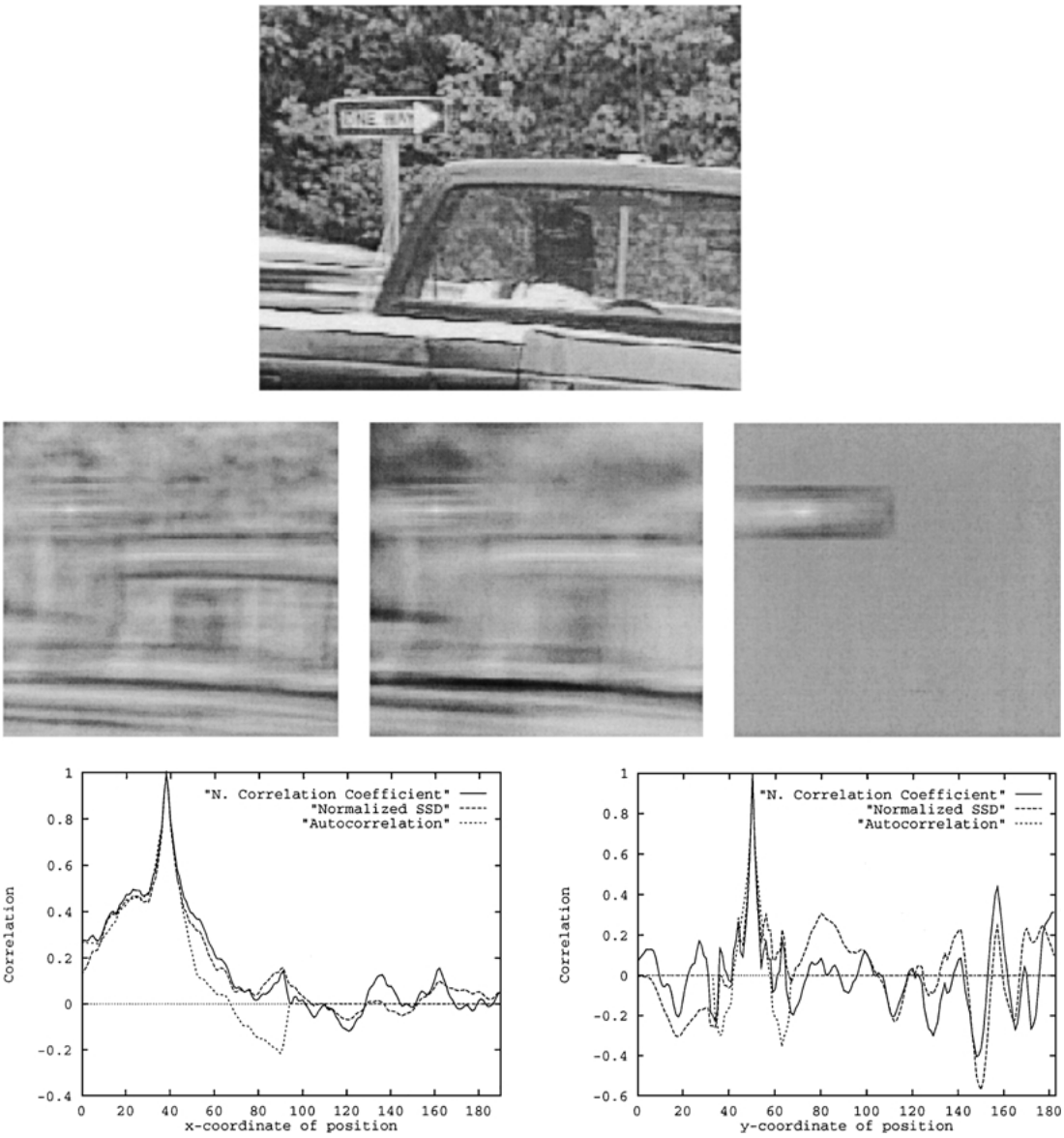


Figure 15. Above, a scene image with a one-way sign. In the middle, three ambiguity surfaces computed for all possible translations of the one-way sign replica with fixed angle and scaling parameters. The left surface is computed using the normalized correlation coefficient, the middle surface is computed using the normalized sum-squared difference, and the right surface is the autocorrelation. The correlation peak of the surfaces is a white spot located in the upper left of each plot. Below, horizontal and vertical slices through the global peaks of the ambiguity surfaces. The left graph shows slices along the x-axis of the ambiguity surfaces with the y-coordinate fixed, and the right graph shows slices along the y-axis with the x-coordinate fixed. The methods converge at the true solution.

bound is equivalent to the minimum variance estimator. In the vicinity of the true value of the parameter vector describing the object, the normalized correlation coefficient then behaves as a minimum variance estimator.

The ambiguity surfaces in Fig. 15 contain numerous local maxima, illustrating the fact that object recognition is an inherently nonlinear problem that requires a global optimization procedure for its solution. A brute-force solution, for example, would be an exhaustive search for the global maximum of the normalized correlation coefficient's ambiguity surface. In high SNR, and for high-complexity objects, we find that this approach is robust, as we will describe in

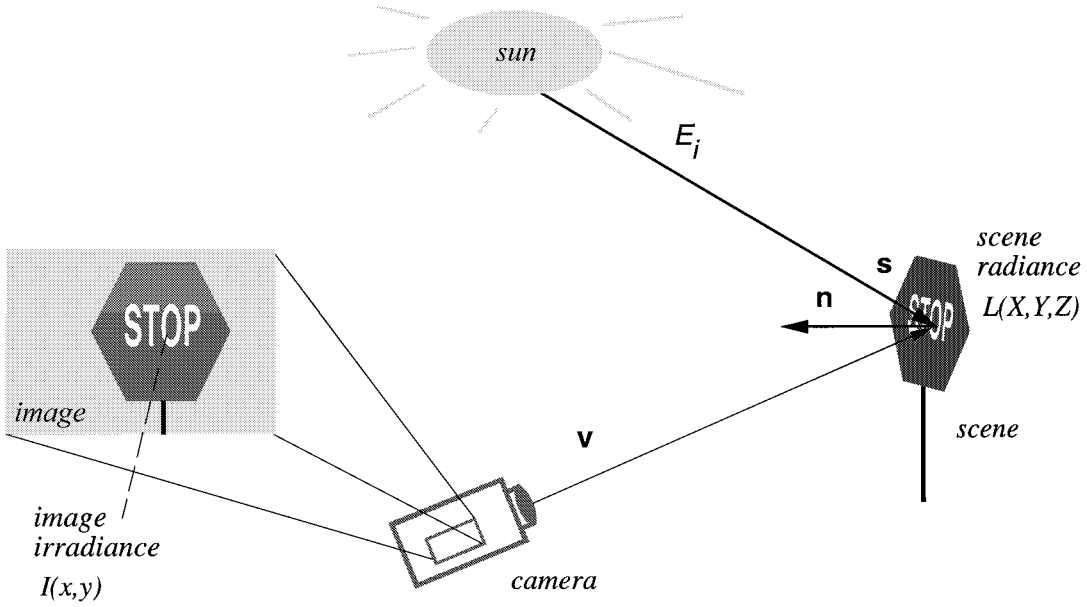


Figure 16. The image irradiance $I(x,\ y)$ is a function of the corresponding scene radiance $L(X,\ Y,\ Z)$, which depends on the source direction **s**, the viewer direction **v**, the surface normal **n**, and the scene irradiance E_i.

Section 15.1. Searching for objects with low complexities, however, may result in false matches. A false match occurs when the object to be recognized has an ambiguously high correlation with a portion of the background scene. This type of ambiguity is typically referred to as "clutter" in radar and sonar detection and is often described as a *deterministic* phenomenon. Classical estimation theory offers no solution to the ambiguity problem since it is based solely upon *statistical* optimality criteria. To prevent mismatches due to deterministic ambiguities, we find that a computer vision system must enforce sufficiently large background and template complexities, as is described in Section 15.1.

9. 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

$$W(x,\ y) = gL(X,\ Y,\ Z), \tag{41}$$

where g is a function of parameters of the imaging system (Horn, 1986). Since the sensitivity of our imaging system is uniform over the whole image, we can assume that g is constant. The imaging scenario is illustrated in Fig. 16.

The scene radiance is related to the object's bidirectional reflectance distribution function (BRDF) f_r and the source irradiance E_i by

$$\begin{aligned} L_r(X, Y, Z) \ = \ & f_r(\mathbf{s}(X, Y, Z), \mathbf{v}(X, Y, Z), X, Y, Z) \\ & \times E_i(\mathbf{s}(X, Y, Z)), \end{aligned} \tag{42}$$

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 $\mathrm{s} = \mathrm{s}(X, Y, Z)$. Under the benign assumption that the object's reflectance has directional properties that are separable from its spatial properties, we have

$$\begin{aligned} f_r\ (\mathbf{s}, \mathbf{v}(X, Y, Z),\ X, Y, Z) \ = \ & f_{r1}(\mathbf{s}, \mathbf{v}(X, Y, Z)) \\ & \times \varrho(X, Y, Z), \end{aligned} \tag{43}$$

where $\varrho(X, Y, Z)$ 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

$$I = f_{r2}(\mathbf{s}, \mathbf{v})\, \varrho^{-\gamma}(X, Y, Z), \tag{44}$$

which, to within the constant factor

$$f_{r2}(\mathbf{s}, \mathbf{v}) = (g f_{r1}(\mathbf{s}, \mathbf{v})\, E_i(\mathbf{s}))^{-\gamma}, \tag{45}$$

is invariant to changes in the geometry of the source, receiver and object.

By distributivity, these results are easily extended to a continuity of sources over a hemisphere, 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 $f_{r2}(\mathbf{s},\ \mathbf{v})$.

Under the relatively benign condition of separability, given in Eq. (43), the resulting Eq. (44) then generally applies to flat surfaces, not too near to the camera, regardless of their reflectance properties. In the case of a Lambertian surface, where $f_{r2}(\mathbf{s},\ \mathbf{v}) = 1/\pi$, however, the result is also valid regardless of whether $\mathbf{v}(X, Y, Z)$ is effectively constant or not, which means that it also applies to the case that the camera is very close to the object.

10. Recognition of Flat Objects

The normalized correlation coefficient, given in Eq. (40), is invariant to linear transformations of image brightness of the form

$$I'(x, y) = c_1 I(x, y) + c_2, \tag{46}$$

where c_1 and c_2 are scalar constants as shown in Appendix C. 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 normalized correlation coefficient, therefore, is invariant to such changes in scene shading, as is our optimal estimate of the parameters **a**, and as is necessary for object recognition.

Uncooperative conditions such as strong shadows and occlusion can change the image irradiance nonuniformly and will cause problems with recognition if they are not accounted for. However, this is not an exclusive weakness of our present formulation, since any approach, for example, systems based on contour or edge detection, will have difficulties in such unpredictable and adverse situations.

11. The Traffic Sign Recognition System

Our method's performance has been evaluated experimentally by applying it to the problem of recognizing traffic signs. This application is very valuable for intelligent vehicles, which can use the recognition results to adjust their speeds or localize themselves in their environments (Betke and Gurvits, 1997). A survey of related papers on traffic sign recognition can be found in Betke and Makris (1997). Our first results were published in Betke and Makris (1995, 1998). Our method stands apart from previous approaches, because it is not restricted to edge detection and does not rely on color information. In principle, our approach could be extended by parameterizing color information.

An overview of our traffic sign recognition system is shown in Fig. 17. The system has a library of replica models, one for each traffic sign class, which are input along with a scene image. It outputs a description **a** of the recognized traffic sign in the scene image or concludes that the scene does not contain a traffic sign. The system consists of three components: a replica generator discussed in Section 12, an estimator based on normalized correlation as discussed in Section 8, and a parameter perturbation component discussed in Section 13.

The recognition process starts by choosing an arbitrary model class ν and an initial parameter vector **a** randomly. The replica generator uses the initial guess of **a** to transform the model image into replica image $q(x, y; \mathbf{a}, \nu)$. The normalized correlation coefficient $r(\mathbf{a})$ is used to evaluate the match of the replica with the scene. If the match is poor, meaning it does not satisfy a predetermined threshold δ on r, the parameter vector **a** is perturbed using simulated annealing, a standard nonlinear optimization method. With the perturbed parameter vector, a new replica is created and tested.

The process of perturbing and evaluating **a** is iterated for a fixed amount of time and then repeated for all sign classes. If the best-matching replica image $q(x, y; \mathbf{a}^*, \nu^*)$ among all parameters **a** and classes ν correlates highly with the scene image, i.e. $r \geq \delta$, the object is recognized, and the system outputs $\mathbf{a}^*$ and ν^*. Otherwise, the system outputs that no traffic sign was found in the scene.

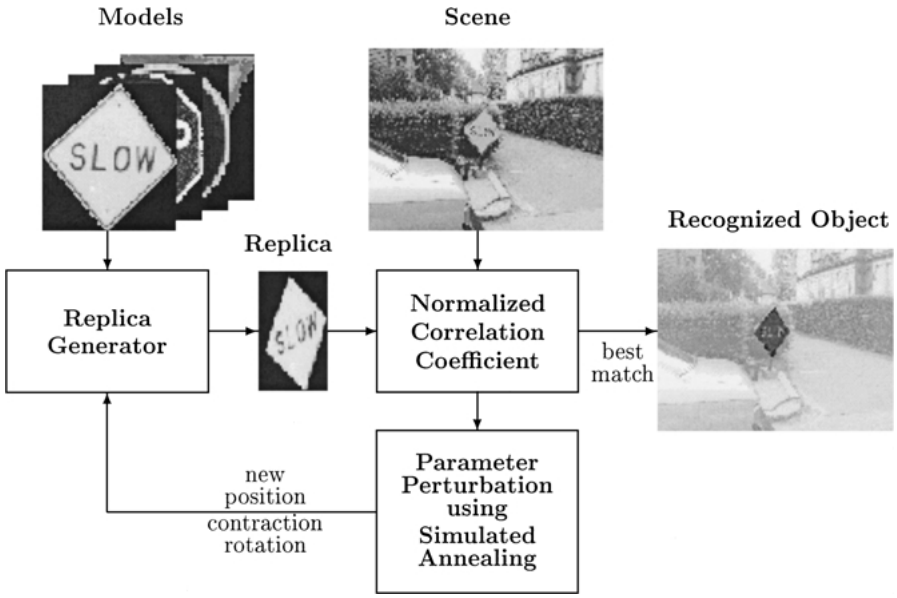


Figure 17. The traffic sign recognition system.

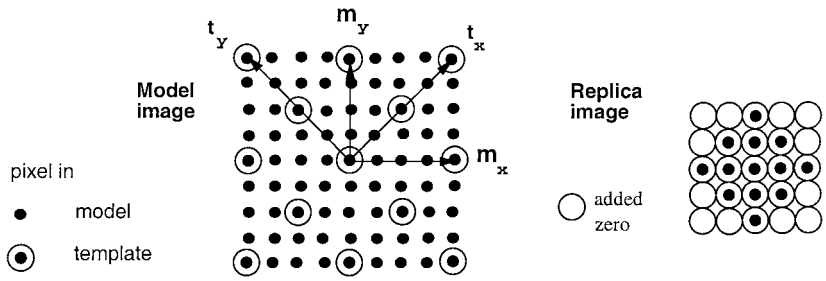


Figure 18. A 5×5 replica image is obtained from a 9×9 model image using contraction parameters $s_x = s_y = 2$ and rotation parameter $\theta_0 = 45°$. For the transformation, the pixels along diagonal vectors $\mathbf{t_x}$ and $\mathbf{t_y}$ are rotated by 45° and become aligned with the coordinate system of the replica. Similarly, the pixels along vectors $\mathbf{m_x}$ and $\mathbf{m_y}$ are rotated by 45° and become aligned with the diagonals of the replica. The replica consists only of the circled pixels of the model and additional zero-brightness pixels where necessary to assure a rectangular image shape.

12. Generating Replicas From Model Images

For efficiency reasons, we use five parameters $(x_0, y_0, \theta_0, s_x, s_y)$ to approximate the affine transformation defined in Eq. (6). The skew parameter α is not varied, but set to zero. This is a valid approximation of transformations that traffic sign images undergo, because the signs are generally fronto-parallel to the image plane or tilted by not more than 45° and are far away from the camera compared to their sizes.

Figure 18 shows how a replica image is generated from a model by subsampling. In this example, the parameters are chosen so that the replica consists only of the circled pixels of the model and additional zero-brightness pixels. In general, a four-point interpolation is used to compute the brightness values of the subsampled replica. Examples of other replicas created by this method are shown in Fig. 19. Our method computes the replica very quickly by sweeping over the model image only once. The time for creating a replica image from a $n \times m$ model image is $O(nm)$.

13. The Simulated Annealing Algorithm

Since the space of possible solutions of the recognition problem is extremely large, the recognition method described here is based on simulated annealing, a popular search technique for solving nonlinear

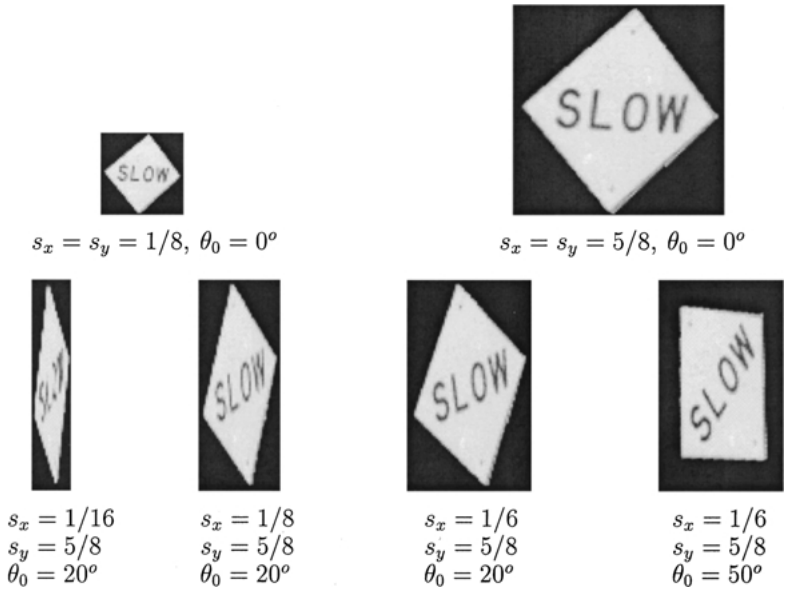


Figure 19. Six replica images of a slow sign, obtained by sampling the slow sign model at various sampling rates and degrees of rotation.

optimization problems (Metropolis et al., 1953; Kirkpatrick et al., 1983), which has been applied to many computer vision problems, e.g., Friedland and Rosenfeld (1991). Its name originates from the process of slowly cooling molecules to form a perfect crystal. The analogue to this cooling process is an iterative search process, controlled by a decreasing "temperature" parameter. At each iteration j, the algorithm generates a replica $q(x, y;\ \mathbf{a}, \nu)$ as described in Section 12. A new test value $a_{\text{test}}^{(j)}$ for parameter a at iteration j is created by

$$a_{\text{test}}^{(j)} = a^{(j-1)} + \Delta a^{(j)}, \tag{47}$$

where $a^{(j-1)}$ is the previous value of a and step $\Delta a^{(j)}$ is a random variable that is uniformly distributed within some interval $[-A,\ A]$. The step bound A is determined experimentally. To properly deal with image boundaries of a scene image, the j-th test value for the center (x_0, y_0) of a replica of width w_x and height w_y is computed by

$$\begin{aligned} x_{0,\text{test}}^{(j)} &= \left(x_0^{(j-1)} + \Delta x_0^{(j)} - w_x^{(j-1)}\right) \\ &\quad \times \text{mod}\left(m_I - w_x^{(j-1)}\right) + w_x^{(j-1)} \\ y_{0,\text{test}}^{(j)} &= \left(y_0^{(j-1)} + \Delta y_0^{(j)} - w_y^{(j-1)}\right) \\ &\quad \times \text{mod}\left(n_I - w_y^{(j-1)}\right) + w_y^{(j-1)}. \end{aligned} \tag{48}$$

This definition avoids "attracting" the replica to the rim or corners of the scene image during the search.

At each iteration, the test values for the rotation and contraction parameters are used to create a replica image and correlate it with the scene image at the test location. If the normalized correlation coefficient $r_{\text{test}}^{(j)}$ increases over the previous coefficient $r^{(j-1)}$, the test parameter values are accepted since a better match is found, i.e.,

$$\text{if} \quad r_{\text{test}}^{(j)} \geq r^{(j-1)} \quad \text{then} \quad a^{(j)} := a_{\text{test}}^{(j)} \tag{49}$$

for each parameter a. If the current match is worse than the previous match, i.e., $r_{\text{test}}^{(j)} < r^{(j-1)}$, the test values are accepted if

$$\exp\left(-\frac{r^{(j-1)} - r_{test}^{(j)}}{T^{(j)}}\right) > \xi, \tag{50}$$

where ξ is randomly chosen to be in [0, 1], $T^{(j)}$ is the temperature parameter in the j-th iteration, and the negative exponent corresponds to the Boltzmann distribution for thermal equilibrium. For a sufficient temperature, this allows "jumps" out of local maxima. The cooling schedule for the j-th update of the temperature parameter is

$$T^{(j)} = T_0/j \qquad \text{for} \quad 1 \leq j \leq L, \tag{51}$$

RECOGNITION ALGORITHM (Scene image I, set of models $\mathcal{M}$)

1. Initialize recognition threshold δ.
2. Initialize parameter domains.
3. Initialize step bounds A_{x_0}, A_{y_0}, A_{s_x}, A_{s_y}, and A_θ.
4. **For** each model class $\nu \in \mathcal{M}$ **do**
5. Initialize position, contraction, and rotation of replica, e.g., at random.
6. Initialize search length L and temperature T_0.
7. **For** $j = 1$ to L **do**
8. Update temperature $T^{(j)} := T_0/j$.
9. Pick $x_{0,test}^{(j)}$ and $y_{0,test}^{(j)}$ randomly according to Eqs. 47 and 48.
10. Evaluate correlation $r_{test}^{(j)}$ as a function of
11. $x_{0,test}^{(j)}$, $x_{0,test}^{(j)}$, $s_x^{(j-1)}$, $s_y^{(j-1)}$ and $\theta_0^{(j-1)}$ according to Eq. 40.
12. Choose ξ uniformly at random within $[0, 1]$.
13. **If** $\exp(-(r^{(j-1)} - r_{test}^{(j)})/T^{(j)}) > \xi$
14. **then** $x_0^{(j)} := x_{0,test}^{(j)}$, $y_0^{(j)} := y_{0,test}^{(j)}$, $r^{(j)} := r_{test}^{(j)}$, update best replica q_ν^*
15. **else** $x_0^{(j)} := x_0^{(j-1)}$, $y_0^{(j)} := y_0^{(j-1)}$, $r^{(j)} := r^{(j-1)}$.
16. Pick $s_{x,test}^{(j)}$, $s_{y,test}^{(j)}$ and $\theta_{0,test}^{(j)}$ randomly according to Eq. 47
17. Create new replica with contractions $s_{x,test}^{(j)}, s_{y,test}^{(j)}$ and rotation $\theta_{0,test}^{(j)}$.
18. Evaluate correlation $r_{test}^{(j)}$ as a function of
19. $x_0^{(j)}$, $y_0^{(j)}$, $s_{x,test}^{(j)}$, $s_{y,test}^{(j)}$ and $\theta_{0,test}^{(j)}$ according to Eq. 40.
20. Choose ξ uniformly at random within $[0, 1]$.
21. **If** $\exp(-(r^{(j)} - r_{test}^{(j)})/T^{(j)}) > \xi$
22. **then** $s_x^{(j)} := s_{x,test}^{(j)}$, $s_y^{(j)} := s_{y,test}^{(j)}$, $\theta_0^{(j)} := \theta_{0,test}^{(j)}$, $r^{(j)} := r_{test}^{(j)}$,
23. update best replica q_ν^*.
24. Optimize q_ν^* by small local parameter perturbations.
25. Determine replica q^* with highest correlation among all q_ν^*.
26. **If** correlation for $q^* < \delta$
27. **then** output "*No traffic sign found in image I.*"
28. **else** output "*Traffic sign $q(x, y; x_0^*, y_0^*, \theta_0^*, s_x^*, s_y^*; \nu^*)$ found.*"

Figure 20. The recognition algorithm.

where T_0 is the initial temperature and L is the number of iterations during the search. Equation (51) describes the fast converging inverse linear cooling schedule (Szu and Hartley, 1987). See Strenski and Kirkpatrick (1991) for a thorough comparison of annealing algorithms with finite length cooling schedules. Since, after L iterations, the search may not have yielded the optimal solution, a local exhaustive search is conducted around the best solution found. The best result of the local search among all classes describes the recognized sign, as long as it has a normalized correlation coefficient that lies above threshold δ. The pseudo code of the recognition algorithm is shown in Fig. 20. The behavior of the parameters during a typical run of the algorithm is shown in Fig. 21.

14. The Computational Complexity of the Search

The number of possible solutions of the recognition problem depends on the number of possible parameter values for traffic sign class, position, rotation, and contraction. For a typical image of size 320×240, the number of possible positions is 76,800. For a number of possible contractions in x and y of both 30, a number of possible rotations of 20, and a class size of 9, the full search space has a size of 1.2×10^{10}. Evaluating the full search space exhaustively is too slow to be practical.

Although a comparison of the simulated annealing algorithm with an exhaustive search may seem unfair, it is nevertheless instructive. The annealing algorithm finds a traffic sign in less than 7000 iterations. That means that only 7000 possible solutions instead of 1.2×10^{10} are evaluated, which is a speedup of several orders of magnitudes over an exhaustive search.

Figure 21 reports a typical run of the algorithm, which takes ca. 100 seconds per sign on a 333 MHz PC running Linux.

15. Experimental Results

Our data consists of more than 3280 scene images, a few of which are shown in Fig. 22. The main criterion for the selection of the scene images is to obtain a wide variety of traffic sign scenes, originating from both the U.S. and Europe. The signs in the scenes have

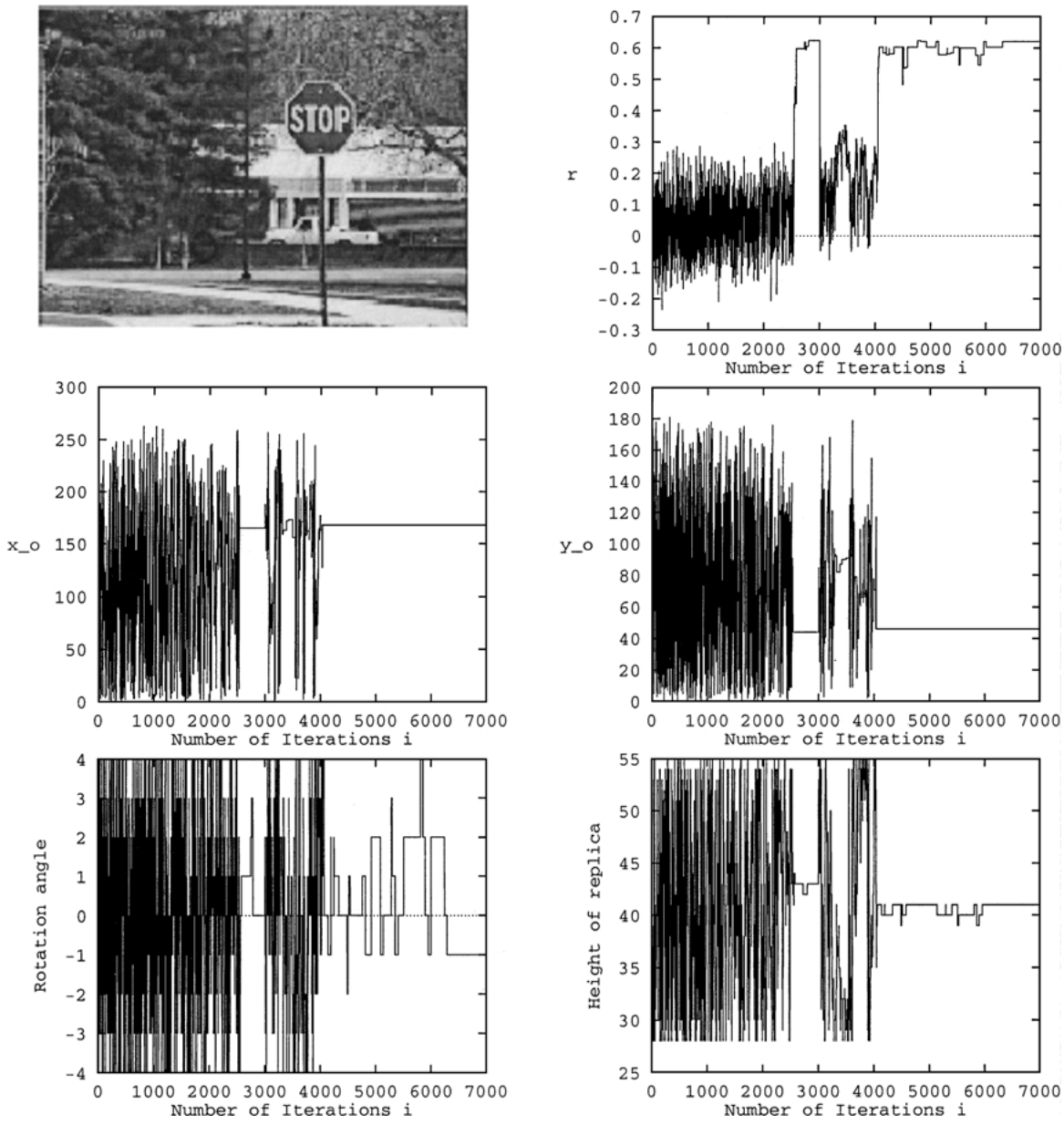


Figure 21. Five graphs illustrating the behavior of the correlation coefficient r and the parameters x_0, y_0, θ_0 and s_y during a typical run of the simulated annealing algorithm with a stop sign scene as input. The algorithm is run with the initial parameters reported in Fig. 24. The algorithm takes ca. 2 *min* per sign on a 333 MHz PC running Linux. The stop sign is almost recognized at iteration 2802, but the temperature parameter is still too high for parameter convergence. The sign is finally recognized at iteration 4050, after which the parameters are only slightly adjusted.

different sizes and orientations, are illuminated differently, and have various backgrounds. Some traffic signs are aged and bent, some are painted with graffiti. Some street scenes do not contain any traffic signs. The model images used to represent the traffic sign classes are shown in Fig. 9. The model traffic signs are physically different signs from the signs in the test scenes. One hundred training images are used to determine the recognition threshold.

The performance of our traffic sign recognition system depends on the complexities of the signs in the images. The system recognizes 94% of the traffic

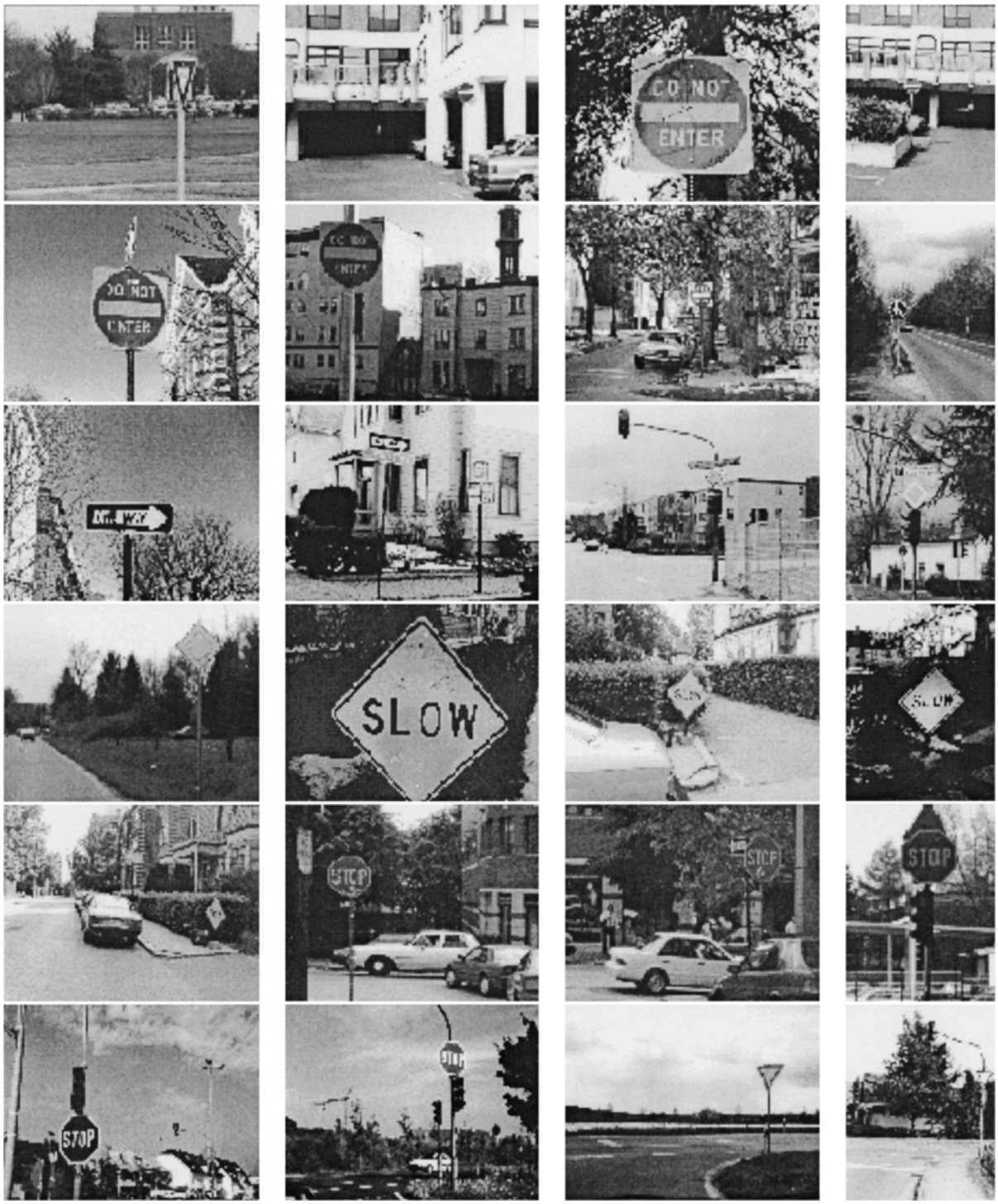


Figure 22. Some of the images used in the recognition experiments.

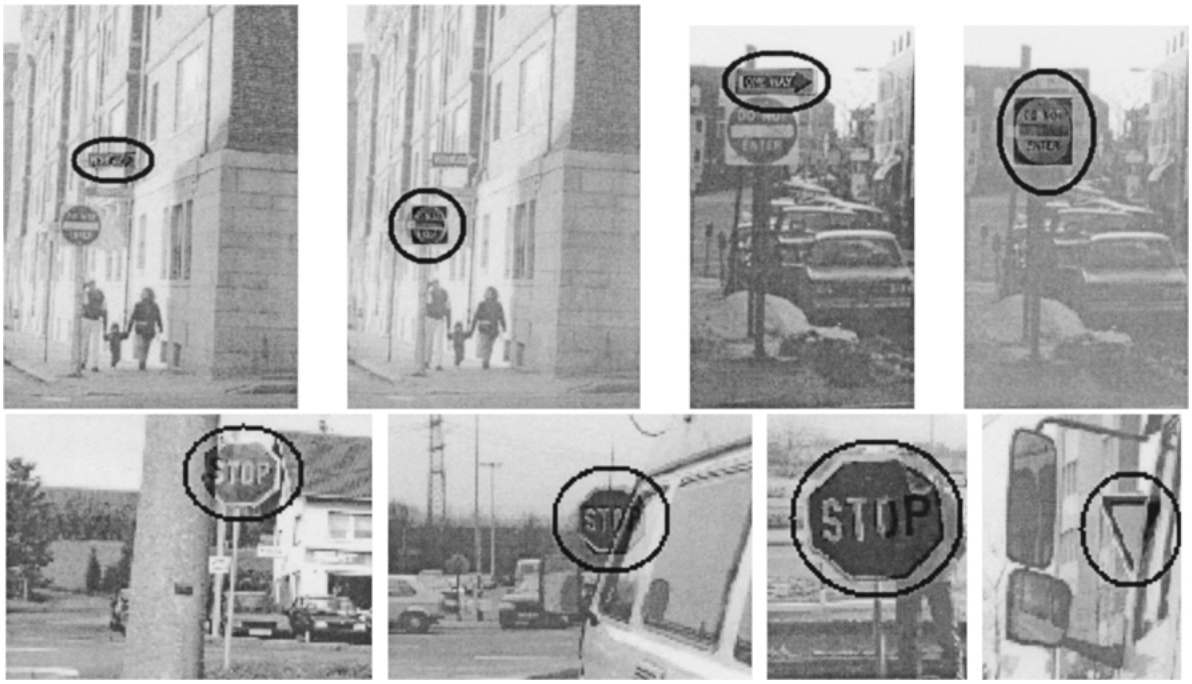


Figure 23. Recognizing multiple and occluded signs: the best matching replica is shown overlying the circled sign in the image.

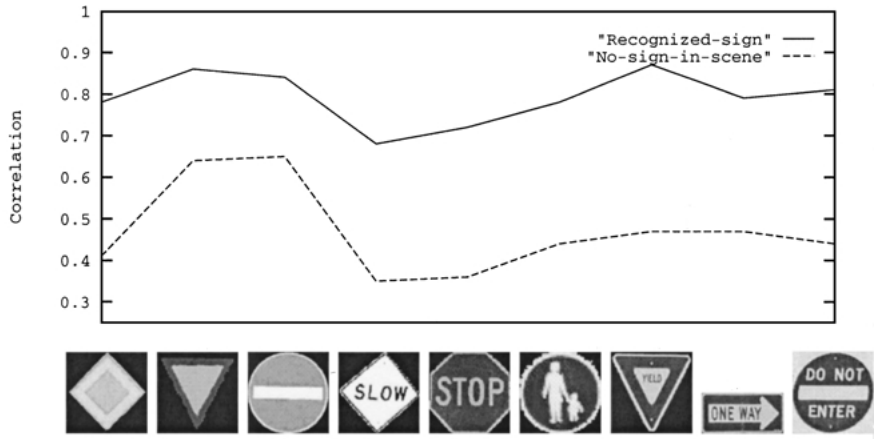
signs correctly and misclassifies 6%, *provided that the positional complexity of the template signs is not smaller than 37*. These numbers discount mismatches of European signs with their corresponding American signs. For example, replicas of European yield signs do not have any inscriptions, but correlate highly with scenes of American signs with the inscription "yield." Figure 23 shows some recognition results, including scenes with occluded signs and with several traffic signs. (The pseudo-code in Fig. 20 is easily modified in line 25 so that several signs in a scene can be found). The initial parameter values, step and domain bounds used in the simulated annealing algorithm are listed in Fig. 24.

Recognition threshold:	$\delta = 0.6$
Search length:	$L = 7000$
Initial temperature:	$T_0 = 210$
Initial position:	$(x_0^{(0)}, y_0^{(0)})$ = center of scene
Initial rotation:	$\theta_0^{(0)} = -4^o$
Initial contraction:	replica of size 70×70 pixels
Location domain:	$(x_0^{(j)}, y_0^{(j)})$ anywhere in scene
Rotation domain:	$\theta_0^{(j)} \in [-10^o, 10^o]$
Contraction domain:	positional complexity > 37
Step bounds:	$A_{x_0} = A_{y_0} = 50$ pixels, $A_{\theta_0} = 5^o$, contraction bounds allow ± 10 pixel steps
Local exhaustive search:	position shift by ± 2 pixels replica contraction by ± 2 pixels replica rotation by $\pm 1^o$

Figure 24. Initial parameter values, step and domain bounds for annealing algorithm.

The solid graph in Fig. 25 illustrates the average best correlation for correct matches. For example, replicas created from the footpath sign model match correctly with scene images that contain footpath signs with an average correlation of 0.78. An average of 408 scene images per traffic sign was used, except for the rare slow sign for which only a few images could be obtained. The dashed graph in Fig. 25 illustrates the average best correlation for scene images that do not contain a traffic sign in the correct class or do not contain a traffic sign at all. For example, the best correlation of a stop sign replica with an arbitrary image that does not contain a stop sign is 0.36 on average. The average is taken over about 2200 scene images per sign class. Note for comparison that the average correlation for a replica that matches with an arbitrary scene is zero, $E[r] = 0$, while a perfect match yields $r = 1$.

A comparison between the average best correlation for recognized signs and for scenes without signs shows that the correlation for a correct match is high enough to identify the correct sign class uniquely among the 9 classes. For almost all scene images, the correlation is highest for matches between a sign in the image and its corresponding replica.



Correct matches:	339	336	380	4	503	327	500	504	376
Correct negatives:	2455	1942	2199	2601	2245	2120	1631	2203	2209
False positives:	1	41	119	1	0	0	0	0	0
Positional complexities:	39	179	199	224	652	835	919	1199	1337

Figure 25. Recognition results for the nine model signs, ordered by their complexities as in Fig. 9. The solid graph plots the average best correlation for scenes with a given sign recognized, and the dashed graph plots the average best false correlation for scenes without a given sign. The comparison shows that the correlation for a correct match is high enough to identify the correct class uniquely among the 9 classes, because the false positive correlations are always lower on average. Underneath the sign models, a table lists the number of correct positive matches and, for scenes without corresponding signs, the number of correct negative and false positive matches. The sum of the entries in each column gives the total number of scene images used for each model class. The internal positional complexities of the signs are reported from Fig. 6 illustrating that false positive matches only occur for low-complexity signs. For signs with sufficiently high complexity, only true positive matches occur, and therefore the Cramer-Rao lower bound on position estimation error is attained.

False positive matches occur when the best correlating replica is not in the correct class. Figure 26 shows an example of a false positive match where the no-entry sign to be recognized is covered by graffiti and occluded by a nonuniform shadow. Although the sign in the scene can be found, as shown in the left image in Fig. 26, the corresponding match yields a normalized correlation coefficient that is slightly lower than the coefficient due to the best-matching European yield sign, as shown in the right image in Fig. 26. As can be seen from the data in Fig. 25, the European no-entry and European yield models generally result in high normalized correlation coefficients for arbitrary scenes and are therefore responsible for the vast majority of false matches.

15.1. The Impact of Object Complexity on Recognition Ambiguity

The problem of ambiguity is one that any object recognition system must eventually address in practice. Image ambiguities arise, for example, when a portion of the background scene has high coincidental correlation with the scene object to be recognized, as illustrated in Fig. 4 of Section 2. Ambiguities can then occur in the absence of noise and are somewhat analogous to background *clutter* in the radar/sonar problem (Difranco and Rubin, 1968).

In our experiments, we find that the number of ambiguities encountered depends on the complexities of the model image and the template created from this model, which are independent of the data, as well as the complexities of the scene object itself and the unoccluded background of the correlation overlap window. When the class of the template and scene object are identical, the recognition system finds a correct match with high probability so long as both template and scene object complexities are high. When low-complexity templates or scene objects are involved, however, recognition ambiguities occur at high frequency. We find that the number of these ambiguities falls off exponentially as complexity increases and exploit this inverse

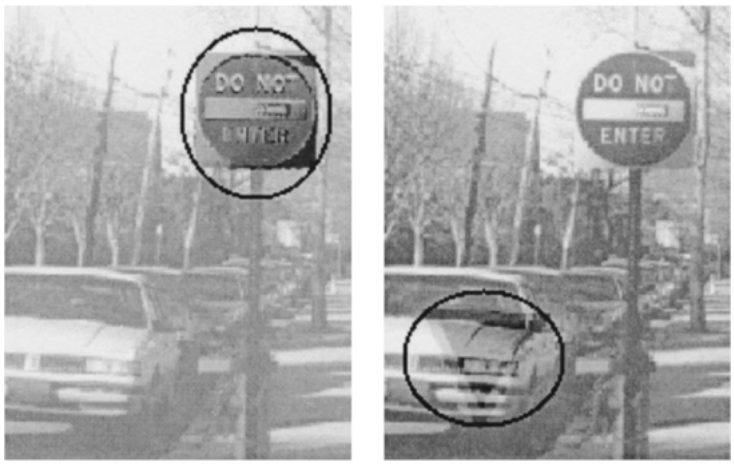


Figure 26. Results for an image with a nonuniform shadow: On the left, the final no-entry template is shown overlying the no-entry sign. On the right, the final yield template is shown matching with the scene background.

relationship between complexity and ambiguity level to enhance system performance.

The impact of *model complexity* on recognition is partially exhibited in Fig. 25 of the previous section where false positive matches occur exclusively for low-complexity models while high-complexity models are effectively immune to ambiguity. This successful performance is primarily achieved by our preconditioning of the lower threshold allowable for template complexity. Without such preconditioning, even high complexity models would suffer significant recognition ambiguity after downsampling, which in turn would translate to a much higher overall level of false positive matches.

It is then important to understand the impact of *template complexity* on recognition. Templates are just downsampled models, so for example, a minor reduction in the size of a high-complexity model will yield a high-complexity template. Typically, *high-complexity templates* either yield an unambiguously high correlation with the correct scene object, which gives a correct positive match, or a low correlation with the scene background, which gives a correct negative match.

Consider, for example, the high-complexity template of Fig. 27(b) which is unambiguously recognized by our simulated annealing algorithm in the scene of Fig. 27(a). The high level of coregistration between

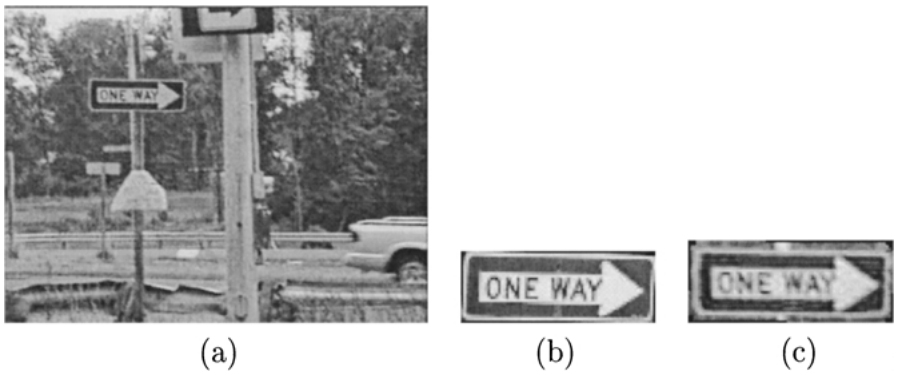


Figure 27. The impact of high template complexity on recognition: The oneway sign in the scene shown in (a) is uniquely recognized with a correlation coefficient of 0.82. (b) The matching template, shown enlarged, has internal complexity 250. (c) The scene sign has internal complexity 119 and outer complexity 140.

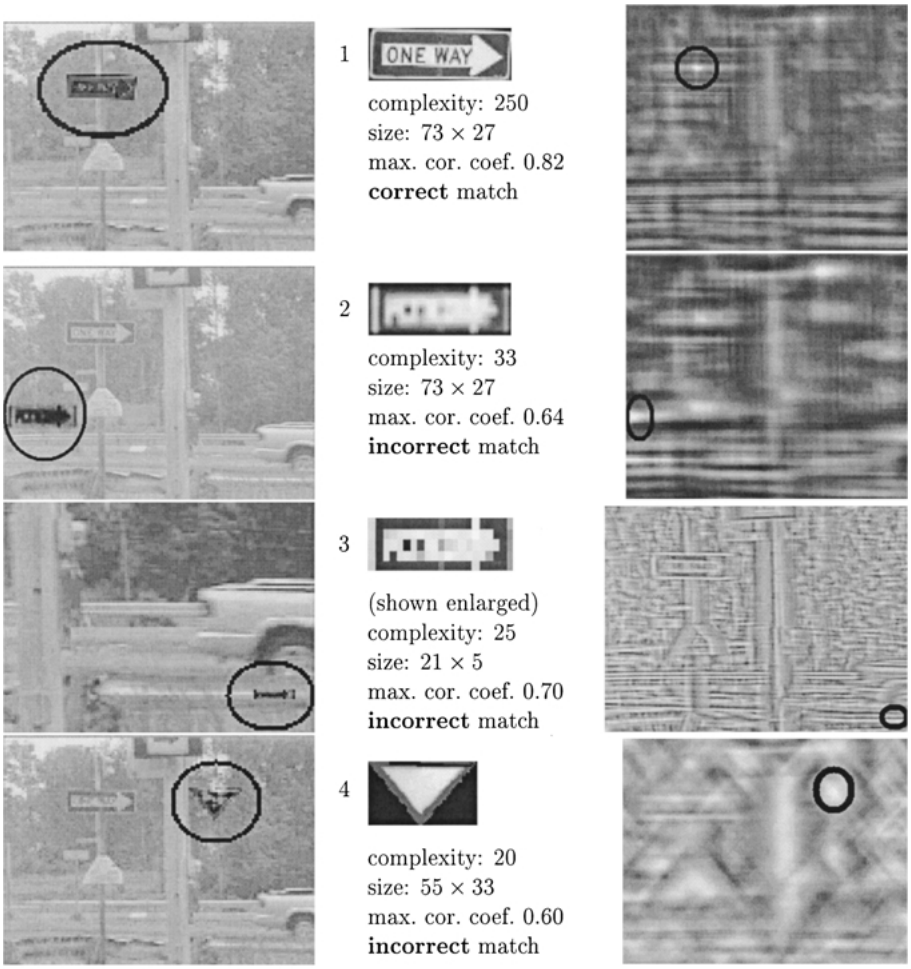


Figure 28. Exhaustive search results for the scene in Fig. 27. Column (a): Four subimages of the scene with overlapping templates at the positions that yield the maximum correlation coefficients. Column (b) lists the internal positional complexities, sizes, and maximum correlation coefficients for the four templates. Column (c): The ambiguity surfaces computed for all possible translations of the four templates. The positions that yield the maximum correlation coefficients are circled. The *low-complexity* templates 2 through 4 highly correlate with the scene background and produce ambiguity surfaces with many spurious peaks. Only the *high-complexity* template 1 correctly matches the oneway sign in the scene.

template and scene object achieved at the correct value of the object parameter vector leads to a well-defined global maximum at the peak of the positional ambiguity surface, as shown in Fig. 28(b)-1. More importantly, however, Fig. 29 shows the maximum correlation coefficient versus template complexity for the correct model object, in this case the oneway sign, as well as other model objects that do not appear in the scene, in an exhaustive search over the scene. While the figure is for a specific example, the behavior it describes is characteristic of that found in a wide range of recognition problems. Specifically, at very low template complexities, i.e. under 35 in this case, all signs have maximum correlations high enough to produce false positive matches. A lower limit on template complexity, i.e. 35 in this case, then must be set to avoid this problem. With increasing template complexity, however, the maximum correlation coefficient falls off rapidly for all models, with one important exception. This exception occurs when the complexity of the template for the correct model approaches the complexity of the scene object. The maximum correlation coefficient then dramatically increases, with large fluctuation, until it reaches a global maximum where the

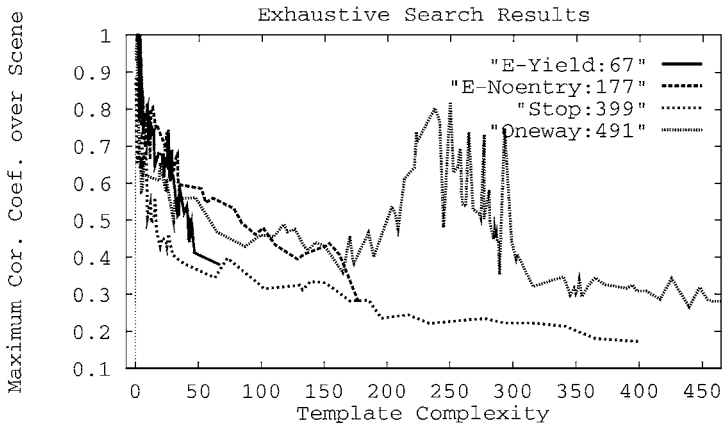


Figure 29. The maximum correlation coefficient in an exhaustive search over the scene shown in Fig.27(a) versus template complexity for European yield and no-entry, stop, and oneway signs, which have internal model complexities of 67, 177, 399, and 491, respectively. Only the oneway sign is in the scene and is recognized with a correlation coefficient of 0.82 at complexity 250. Note the significant number of ambiguities for template complexities below 35.

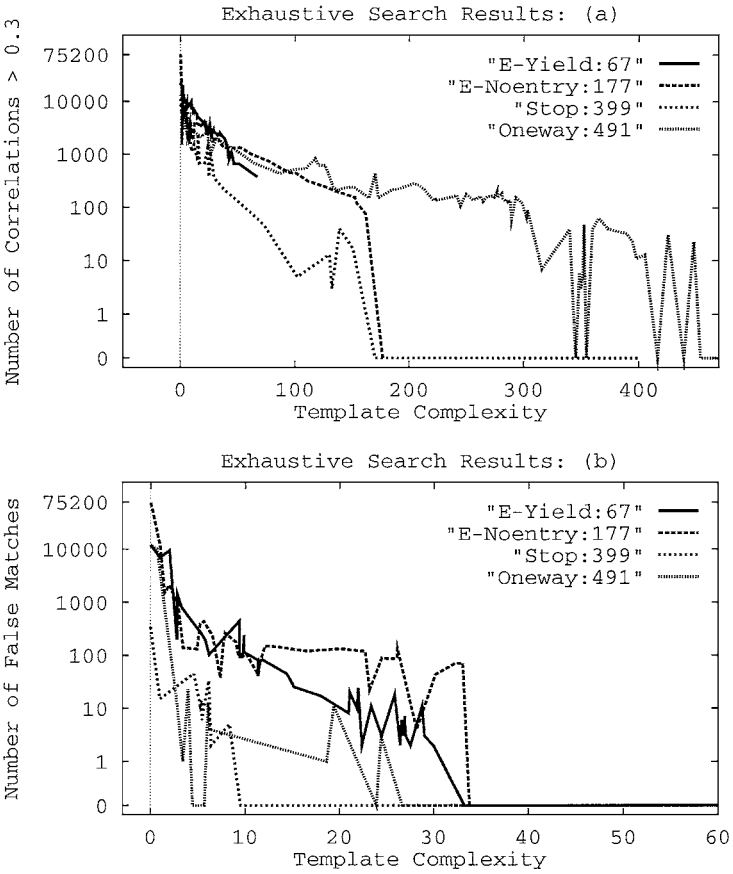


Figure 30. The number of matches versus positional template complexity in an exhaustive search over the scene shown in Fig. 27(a) for European yield and no-entry, stop, and oneway signs, which have internal positional model complexities of 67, 177, 399, and 491, respectively. Only the oneway sign is in the scene. (a) The number of correlation coefficients above 0.3. (b) The number of false matches, which are defined to be correlation coefficients above 0.6. Note the significant number of ambiguities for complexities below 35.

template and scene object complexities coincide. Figure 29 also shows that a high-complexity template belonging to the wrong class yields a low correlation and therefore does not cause a false positive match.

A more in-depth understanding of the relationship between template complexity and recognition ambiguity can be gained by analyzing the total number of correlations above a given threshold, as well as the number of false positive matches, found in an exhaustive search over a particular scene versus template complexity. This is illustrated in Fig. 30, where the most striking behavior is that the number of false positive matches becomes exceedingly large when template complexities are small, regardless of the model. Fortunately, the number of false positive matches falls off exponentially with increasing template complexity until it eventually vanishes at some critical value, i.e. 35 for the given example. The basic issue then is that template complexity must be restricted at the low end to avoid the inevitably large number of false positive matches that would otherwise plague a recognition system. An intuitive understanding of this phenomenon can be obtained by considering the definition of complexity as "the number of coherence cells" contained in the object. Then, as the template complexity decreases, the number of coherence cells also decreases. As the number of coherence cells decreases, the probability that spurious combinations of these cells will constructively interfere with the background to yield an ambiguously high correlation increases.

Low-complexity templates are created when high-complexity models are downsampled. The downsampling is typically necessary to find the correct object scale, but leads to loss of information and a relative decrease in template complexity with respect to the model. To help visualize the problem, an example of

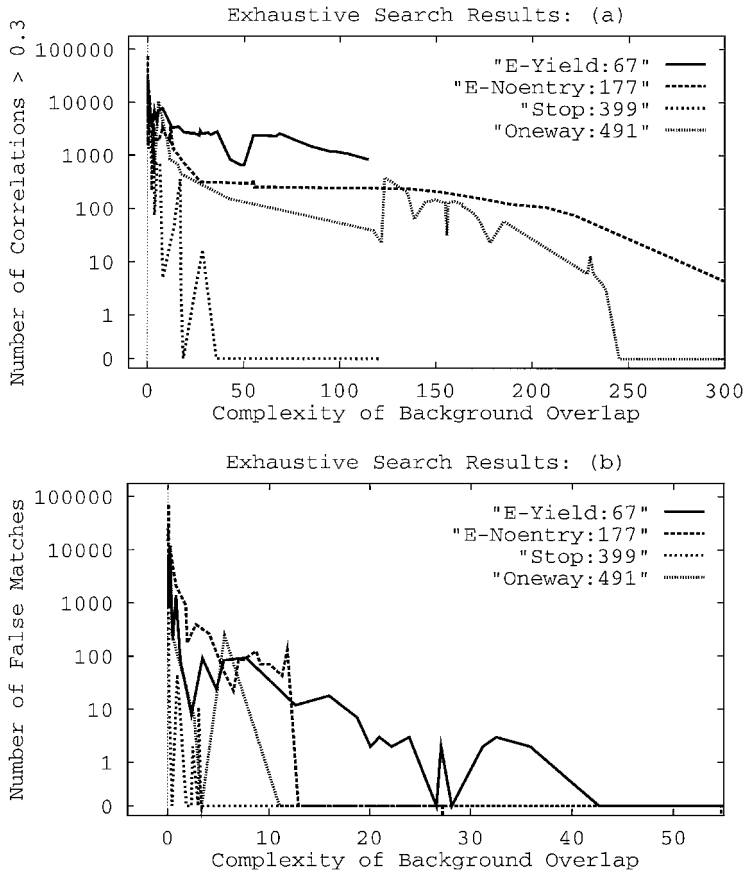


Figure 31. The number of matches versus background complexity in an exhaustive search over the scene shown in Fig. 27(a) for European yield and no-entry, stop, and oneway signs, which have internal model complexities 67, 177, 399, and 491, respectively. Only the oneway sign is in the scene. (a) The number of correlation coefficients above 0.3. (b) The number of false matches, which are defined by correlation coefficients above 0.6.

a low-complexity template that achieves a spuriously high correlation with the scene background is illustrated in Fig. 28(b)-3. Low-complexity templates are also created when low-complexity models are downsampled. In such cases, the relative reduction in complexity may be small, but the final complexity is often so low that false matches occur, as illustrated in Fig. 28(b)-4. Here the template size is roughly half that of the object, but it is of the wrong class, and produces a false positive match.

The complexity of the scene object also significantly impacts the ability of a system to recognize it. Poor resolution or significant occlusion can lead to low complexity in a scene object. If a low-complexity template of the wrong class is then used to identify the object, incorrect matches are likely. Conversely, use of a high-complexity template of the correct class typically leads to a correct positive match. For example, in Fig. 28(b)-1, a high-complexity template correctly matches the lower complexity scene object shown in Fig. 27(c).

Similarly, the complexity of scene background in the overlap window of the correlation has a strong inverse relationship with the level of recognition ambiguity. As both the complexities of the template and overlapping background decrease, the possibility of a spurious false match between the two becomes more likely. This phenomenon is illustrated in Fig. 31, where the number of matches is plotted versus background complexity. Background overlap regions of low complexity yield a large number of ambiguities. The number of ambiguities decreases exponentially with increasing background overlap complexity so long as the template complexity is sufficiently high.

It is significant that the size of a template, scene object or background overlap window is not a good inverse measure of recognition ambiguity. This is illustrated in Fig. 28 where the low-complexity template shown in Fig. 28(b)-3 is scaled up by linear interpolation so that gradients and complexity are preserved to within the discretization error. The resulting template, shown in Fig. 28(b)-2, has the same size as the high-complexity template, shown in Fig. 28(b)-1 but the same complexity as the low-complexity template of Fig. 28(b)-3. The high-complexity template, however, yields a correct match while the low-complexity template of the same size and model yields a false match. This example illustrates the fact that it is the complexity of an template, scene object or background overlap window that is inversely proportional to the recognition ambiguity, not the size. The size only provides an outer scale, but it is the ratio of the inner coherence scale to the outer size scale that is important in assessing the level of ambiguity to be expected.

16. Conclusions

The problem of object recognition for objects subject to affine transformation, including the issue of recognition ambiguity, can be characterized in terms of physical descriptors for object resolution, coherence, and complexity. We have derived analytic expressions for these descriptors, where for example, the resolution of an object subject to affine transformation is given in terms of generalized coherence scales that are extracted directly from a scene's Fisher information matrix. Use of the Fisher information matrix in this manner is novel and advantageous because it enables scalar coherence scales or multi-dimensional coherence volumes to be consistently defined for any set of the affine parameters defining the object. The formulation has the desired limiting behavior since our coherence scale for 1-D position estimation reduces to the coherence length defined in the signal processing literature by Gabor (1946). Our generalized coherence scales are shown to be of great practical value because they correspond to the width of the object's autocorrelation peak under affine transformation and so provide a direct measure of the extent to which an object can be resolved under affine parameterization. The theory shows that resolution increases as the contrast between object and scene background along the object perimeter is increasingly nonuniform.

We then develop a method for recognizing objects subject to affine transformation in complex real world scenes based on template matching. In the search for the best matching template, the affine parameter vector that describes the template is perturbed using simulated annealing, a standard nonlinear optimization procedure. The match is quantitatively evaluated using the normalized correlation coefficient. The method is then used to recognize traffic signs in thousands of real-world scenes.

We show that our measure of complexity, derived in terms of the object's generalized coherence scales, has a strong inverse relationship to the level of recognition ambiguity, whereas other potential metrics, such as the relative size of the scene object or template, exhibit no such simple relationship. We exploit these findings to

reduce the level of recognition ambiguity by preconditioning the permissible range of template complexity above a priori thresholds.

For many three-dimensional objects, the affine vector of the present formulation must be supplemented by further parameters 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. Our approach, however, can be directly applied to this more general problem to derive recognition methods, object coherence scales and complexities so long as partial derivatives of scene brightness can be sensibly defined in terms of the components of the augmented parameter vector, as should be the case when a physical model is employed.

In summary, our analysis shows that object recognition, resolution, coherence, complexity, and ambiguity are all fundamentally related. We have developed practical tools for computing resolution and complexity and demonstrated their importance in the object recognition problem.

Appendix

A. *Signal-Dependent Fluctuations of Natural Light are Negligible in the CCD Brightness Data*

This section shows that thermally induced fluctuations of natural light are not a significant cause of errors in our measurements. Natural light fluctuates as a circular complex Gaussian random (CCGR) process (Goodman, 1985). The probability density of an $M \times N$ intensity image $\mathbf{W}$ measured from a CCGR field is the gamma distribution

$$P(\mathbf{W}) = \prod_{k=1}^{MN} \frac{1}{\Gamma(\mu)} \left(\frac{\mu}{\sigma_k}\right)^{\mu} W_k^{\mu-1} \exp\left(-\mu \frac{W_k}{\sigma_k}\right), \tag{52}$$

where the number of coherence cells μ in the intensity average is defined to be the time-bandwidth product $\mu = T\tau$, where T is the coherence or measurement time, and τ is the bandwidth of the light. The expected value of W_k is σ_k, and the variance of W_k is σ_k^2/μ. Given Eq. (52), the probability density for the "gamma-corrected" brightness $\mathbf{I}$ is

$$P(\mathbf{I}) = \prod_{k=1}^{MN} \frac{1}{\Gamma(\mu)} \left(\frac{\mu}{\sigma_k}\right)^{\mu} I_k^{\gamma\mu-1} \exp\left(-\mu \frac{I_k^{\gamma}}{\sigma_k}\right), \tag{53}$$

since $I_k = W_k^{\frac{1}{\gamma}}$. For notational convenience, the subscript k is dropped in the following. The expected value of I is

$$\mathrm{E}[I] = \int_0^{\infty} W^{\frac{1}{\gamma}} P(W) dW = \left(\frac{\sigma}{\mu}\right)^{\frac{1}{\gamma}} \frac{\Gamma\left(\frac{1}{\gamma}+\mu\right)}{\Gamma(\mu)},$$

and the variance of I is

$$\begin{aligned}\mathrm{var}(I) &= \mathrm{E}[I^2] - \mathrm{E}[I]^2 \\ &= \left(\frac{\sigma}{\mu}\right)^{\frac{2}{\gamma}} \frac{\Gamma(\mu)\Gamma\left(\frac{2}{\gamma}+\mu\right) - \left(\Gamma\left(\frac{1}{\gamma}+\mu\right)\right)^2}{(\Gamma(\mu))^2}.\end{aligned}$$

An approximation of the mean of I using Stirling's Formula that $\Gamma(\mu) = \mu^{\mu-\frac{1}{2}} e^{\mu} (2\pi)^{\frac{1}{2}}$ yields

$$\begin{aligned}\mathrm{E}[I] &\approx \sigma^{\frac{1}{\gamma}} \frac{\left(\frac{1}{\gamma}+\mu\right)^{\frac{1}{\gamma}+\mu-\frac{1}{2}} e^{-\left(\frac{1}{\gamma}+\mu\right)}}{\mu^{\frac{1}{\gamma}}\ \mu^{\mu-\frac{1}{2}} e^{\mu}} \\ &\approx \sigma^{\frac{1}{\gamma}}\ e^{\frac{\frac{1}{\gamma}+\mu-\frac{1}{2}}{\gamma\mu}} e^{-\frac{1}{\gamma}} \approx \sigma^{\frac{1}{\gamma}},\end{aligned}$$

which holds for large μ. The variance of I can be approximated by

$$\mathrm{var}(I) \approx \frac{\sigma^{\frac{2}{\gamma}}}{\mu\gamma^2},$$

and is therefore a function of the mean, which reveals that the noise arising from circular complex Gaussian random fluctuations in the received field is *signal dependent*. This is important for radar and sonar imaging (Makris, 1995), where, due to signal-dependent fluctuation noise, the variance of high-intensity measurements can be larger than the mean of low-intensity measurements. For fluctuations of natural light, however, the intensity average of the measurements is large enough to reduce the standard deviation to a negligibly small fraction of its mean, as shown in the following example for green light.

Green light has a bandwidth of $\tau_{\text{green}} = 3 \times 10^8 \frac{m}{s}$ $(550\ nm - 500\ nm)/(550\ nm \times 500\ nm) = 5.45 \times 10^{13}$ Hz. With exposure time of $T = 1/100$ s, the number of coherence cells is $\mu = 5.45 \times 10^{11}$. The ratio of the standard deviation of I to the mean of I is approximately

$$\frac{\mathrm{std}(I)}{\mathrm{E}[I]} \approx \sqrt{\frac{\sigma^{\frac{2}{\gamma}}}{\mu\gamma^2}}\ \ \frac{1}{\sigma^{\frac{1}{\gamma}}} = \frac{1}{\sqrt{\mu\gamma^2}},$$

which is $O(6\times10^{-7})$, a negligibly small ratio compared to that actually measured in Section 3. Therefore, the

inherent signal-dependent fluctuations of natural light have a negligible effect on our image data.

B. The Lower Bound on Position Estimation

This section analyzes how the lower bound on position estimation, as derived in Section 5.1, varies with changes in object rotation. Let $\mathbf{R}$ be a two-dimensional orthonormal matrix. The Fisher information can then be expressed in terms of $\mathbf{R}$ and a diagonal matrix $\mathbf{D}$ that consists of $\mathbf{R}$'s principal components D_{11} and D_{22},

$$\mathbf{J} = \frac{E}{\sigma^2}\mathbf{B} = \frac{E}{\sigma^2}\mathbf{R}\mathbf{D}\mathbf{R}^T = \frac{E}{\sigma^2}\mathbf{R}\begin{pmatrix} D_{11} & 0 \\ 0 & D_{22} \end{pmatrix}\mathbf{R}^T. \tag{54}$$

where

$$D_{11} = \frac{B_x^2 - B_y^2}{2} + \frac{1}{2}\sqrt{4B_{xy}^4 + \left(B_x^2 - B_y^2\right)^2} \tag{55}$$

$$D_{22} = \frac{B_x^2 - B_y^2}{2} - \frac{1}{2}\sqrt{4B_{xy}^4 + \left(B_x^2 - B_y^2\right)^2}. \tag{56}$$

The angle φ that rotates the x- and y-axes of the image into the principal axes given by the object's bandwidth matrix is defined by

$$\tan(2\varphi) = \frac{2B_{xy}^2}{B_x^2 - B_y^2}. \tag{57}$$

The lower bound on position recognition is therefore

$$\mathbf{J}^{-1} = \frac{\sigma^2}{E}\mathbf{R}\mathbf{D}^{-1}\,\mathbf{R}^T. \tag{58}$$

The lower bound on recognizing the position coordinate x_0 can then be expressed as

$$\mathrm{E}[(\hat{x}_0 - x_0)^2] \geq J_{x_0}^{-1} = \frac{\sigma^2}{E}\frac{D_{11}\,\sin\varphi + D_{22}\,\cos\varphi}{D_{11}\,D_{22}} \tag{59}$$

and on recognizing the coordinate y_0 as

$$\mathrm{E}[(\hat{y}_0 - y_0)^2] \geq J_{y_0}^{-1} = \frac{\sigma^2}{E}\frac{D_{11}\,\cos\varphi + D_{22}\,\sin\varphi}{D_{11}\,D_{22}}. \tag{60}$$

If the coordinate axes correspond to the principal components of the bandwidth of the object, i.e., $B_x^2 = D_{11}$ and $B_y^2 = D_{22}$, then the error in the x- and y-coordinate of the position is lower bounded by $\sigma^2/(EB_x^2)$ and $\sigma^2/(EB_y^2)$, respectively. Let the total estimation error be the Euclidean distance $\xi = \sqrt{\mathrm{E}[(\hat{x}_0 - x_0)^2]^2 + \mathrm{E}[(\hat{y}_0 - y_0)^2]^2}$. The total error is then lower bounded by

$$\xi \geq \frac{\sigma^2}{E}\sqrt{D_{11}^2 + D_{22}^2 + 2D_{11}\,D_{22}\,\sin 2\varphi}. \tag{61}$$

The bound is smallest if the principal components of the object's bandwidth matrix are aligned with the coordinate axes, i.e., $\varphi = 0$ or $\varphi = \pi/2$. The bound is largest for $\varphi = \pi/4$, for which $\mathrm{E}[(\hat{x}_0 - x_0)^2] = \mathrm{E}[(\hat{y}_0 - y_0)^2]$.

C. Linear Invariance of the Normalized Correlation Coefficient

Given the definition in Eq. (40), the normalized correlation coefficient

$$r(I_q,\ c_1 q + c_2) = \frac{A\sum I_q(x,\ y)(c_1 q(x,\ y) + c_2) - \left(\sum I_q(x,\ y)\right)\left(\sum (c_1 q(x,\ y) + c_2)\right)}{\sqrt{A\sum I_q(x,\ y)^2 - \left(\sum I_q(x,\ y)\right)^2}\sqrt{A\sum (c_1 q(x,\ y) + c_2)^2 - \left(\sum (c_1 q(x,\ y) + c_2)\right)^2}}$$

describes how well a linearly transformed replica $c_1 q(x, y) + c_2$ matches with the measured data in subimage $I_q(x, y)$. The numerator of $r(I_q, c_1 q + c_2)$ is

$$c_1\left(A\sum I_q(x,\ y)q(x,\ y) - \sum I_q(x,\ y)\sum q(x,\ y)\right),$$

and the second square root in the denominator is

$$\left(A\sum\left(c_1^2(q(x,\ y))^2 + 2c_1c_2q(x,\ y) + c_2^2\right) - c_1^2\left(\sum q(x,\ y)\right)^2 - 2c_1Ac_2\sum q(x,\ y) - (Ac_2)^2\right)^{1/2},$$

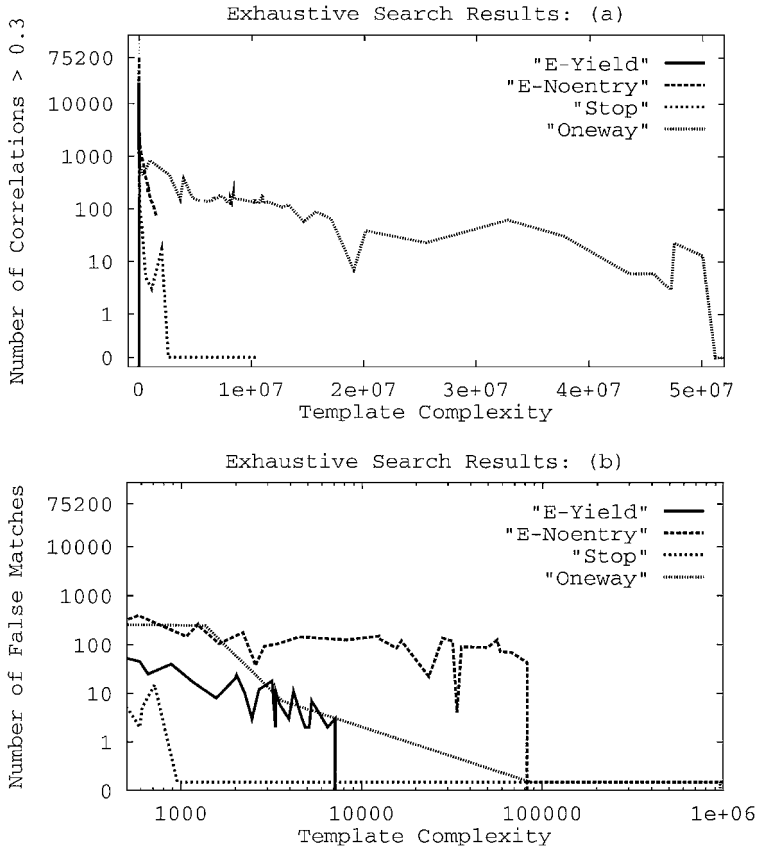


Figure 32. The number of matches versus template complexity in an exhaustive search over the scene shown in Fig.27(a) for European yield and no-entry, stop, and oneway signs, which have internal model complexities of 1.3×10^5, 5.3×10^6, 5.4×10^7, and 8.7×10^7, respectively. Only the oneway sign is in the scene. (a) The number of correlation coefficients above 0.3. (b) The number of false matches occurring for correlation coefficients above 0.6.

which yields $c_1\sqrt{A\sum(q(x,\ y))^2 - (\sum q(x,\ y))^2}$. Since

$$
\begin{aligned}
&r(I_q,\ c_1 q + c_2)\\
&= \frac{c_1\left(A\sum I_q(x,\ y)q(x,\ y) - \left(\sum I_q(x,\ y)\right)\left(\sum q(x,\ y)\right)\right)}{\sqrt{A\sum I_q(x,\ y)^2 - \left(\sum I_q(x,\ y)\right)^2}\ c_1\sqrt{A\sum(q(x,\ y))^2 - \left(\sum q(x,\ y)\right)^2}}\\
&= r(I_q,\ q),
\end{aligned}
$$

the normalized correlation coefficient is invariant to linear transformations of image brightness and Eq. (46) holds.

D. Number of Matches vs. Template Complexity

The impact of positional complexity on recognition was analyzed in Section 15.1. The analysis generalizes to the full complexity that is not just computed for the positional, but also for the rotational and contractional parameters. Figure 32 illustrates the number of matches versus template complexity in an exhaustive search over the scene shown in Fig. 27(a) for European yield and no-entry, stop, and oneway signs, which have internal model complexities of 1.3×10^5, 5.3×10^6, 5.4×10^7, and 8.7×10^7, respectively. Only the oneway sign is in the scene. The figure shows the number of correlation coefficients above 0.3 and the number of false matches occurring for correlation coefficients above 0.6.

References

Ballard, D.H. and Brown, C.M. 1982. *Computer Vision*. Prentice-Hall: Englewood Cliffs, NJ.

Betke, M. and Gurvits, L. 1997. Mobile robot localization using landmarks. *IEEE Trans. Robotics and Automation*, 13:251–263.

Betke, M. and Makris, N.C. 1995. Fast object recognition in noisy images using simulated annealing. In *Proceedings of the Fifth International Conference on Computer Vision*, Cambridge, MA,

June 1995. IEEE Computer Society: Los Alamitos, CA, pp. 523–530.

Betke, M. and Makris, N.C. 1997. Information-conserving object recognition. Technical Report CAR-TR-858, CS-TR-3799, University of Maryland.

Betke, M. and Makris, N.C. 1998. Information-conserving object recognition. In *Proceedings of the Sixth International Conference on Computer Vision*, Mumbai, India, January 1998. IEEE Computer Society: Los Alamitos, CA, pp. 145–152.

Cernuschi-Frias, B., Cooper, D.B., Hung, Y.-P., and Belhumeur, P.N. 1989. Toward a model-based Bayesian theory for estimating and recognizing parameterized 3-D objects using two or more images taken from different positions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11:540–564.

Chin, R.T. and Dyer, C.R. 1986. Model-based recognition in robot vision. *Computing Surveys*, 18(1):67–108.

Difranco, J.V. and Rubin, W.L. 1968. *Radar Detection*. Prentice-Hall: Englewood Cliffs, NJ.

Downie, J.D. and Walkup, J.F. 1994. Optimal correlation filters for images with signal-dependent noise. *Journal of the Optical Society of America A* 11:1599–1609.

Friedland, N.S. and Rosenfeld, A. 1991. Lobed object delineation using a multipolar representation. Technical Report 2779, University of Maryland.

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

Goodman, J.W. 1965. Some effects of target induced scintillation on optical radar performance. In *Proceedings of the IEEE*, 53:1688–1700.

Goodman, J.W. 1985. *Statistical Optics*. Wiley: New York.

Horn, B.K.P. 1986. *Robot Vision*. MIT Press: Cambridge, MA.

Jain, R., Kasturi, R., and Schunk, B. 1995. *Machine Vision*. McGraw Hill: New York.

Kashioka, S., Ejiri, M., and Sakamoto, Y. 1976. A transistor wire-bonding system utilizing multiple local pattern matching techniques. *IEEE Transactions on Systems, Man and Cybernetics*, 6(8):562–569.

Kay, S.M. 1993. *Statistical Signal Processing*. Prentice Hall: Englewood Cliffs.

Kelley, R.B., Martins, H.A.S., Birk, J.R., and Dessimoz, J.-D. 1983. Three vision algorithms for acquiring workpieces from bins. *Proceedings of the IEEE*, 71(7):803–820.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. 1983. Optimization by simulated annealing. *Science*, 220:671–680.

Levanon, N. 1988. *Radar Principles*. John Wiley and Sons: New York.

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

Makris, N.C. 1996. The effect of saturated transmission scintillation on ocean acoustic intensity measurements. *Journal of the Acoustical Society of America*, 100(2):769–783.

Makris, N.C., Avelino, L.Z., and Menis, R. 1995. Deterministic reverberation from ocean ridges. *Journal of the Acoustical Society of America*, 97(6):3546–3574.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equations of state calculations by fast computing machines. 1953. *J. Chem. Physics*, C–21:734–738.

Poynton, C. 1993. "Gamma" and its disguises: The nonlinear mappings of intensity in perception, CRTs, film and video. *J. SMPTE*, 102:1099–1108.

Rao, C.R. 1973. *Linear Statistical Inference and its Applications*, 2nd ed. Wiley and Sons: New York.

Rosenfeld, A. and Kak, A.C. 1982. *Digital Picture Processing*, 2nd ed. Academic Press: New York, Vol. 2.

Strang, G. 1976. *Linear Algebra and its Applications*. Academic Press: San Diego.

Strenski, P.N. and Kirkpatrick, S. 1991. Analysis of finite length annealing schedules. *Algorithmica*, 6:346–366.

Szu, H. and Hartley, R. 1987. Fast simulated annealing. *Physics Letters A*, 122:157–162.

Trucco, E. and Verri, A. 1998. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall: New York.

Umbaugh, S. 1998. *Computer Vision and Image Processing*. Prentice Hall: Englewood Cliffs, NJ.

Van, H.L. Trees. 1968. *Detection, Estimation, and Modulation Theory*, Part I. Wiley: New York.

Yoshimura, S. and Kanade, T. 1994. Fast template matching based on the normalized correlation by using multiresolution eigenimages. In *Proceedings of the International Conference on Intelligent Robots and Systems*, Munich, Germany, September 1994.

Zadeh, L.A. and Ragazzini, J.R. 1952. Optimum filters for the detection of signals in noise. In *Proc. IRE*, 40:1123–1131.