SRI International

# OBJECTIVE FUNCTIONS FOR FEATURE DISCRIMINATION

Technical Note No. 465

May 1989

By: Pascal Fua and Andrew J. Hanson

Artificial Intelligence Center
Computer and Information Sciences Division

The research reported herein was supported by DARPA Contracts MDA903-83-C-0027 and DACA76-85-C-0004.



SRI International®

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334-486

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **MAY 1989** | | **00-05-1989 to 00-05-1989** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Objective Functions for Feature Discrimination** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT **unclassified** | b. ABSTRACT **unclassified** | c. THIS PAGE **unclassified** | | **8** | |

# Objective Functions for Feature Discrimination

Pascal Fua and Andrew J. Hanson*
Artificial Intelligence Center
SRI International
333 Ravenswood Avenue
Menlo Park, California 94025

## Abstract

We propose and evaluate a class of objective functions that rank hypotheses for feature labels. Our approach takes into account the representation cost and quality of the shapes themselves, and balances the geometric requirements against the photometric evidence. This balance is essential for any system using underconstrained or generic feature models. We introduce examples of specific models allowing the actual computation of the terms in the objective function, and show how this framework leads naturally to control parameters that have a clear semantic meaning. We illustrate the properties of our objective functions on synthetic and real images.

## 1 Introduction

All approaches to the problem of extracting features from images can in principle be phrased in terms of decision theory; however, the concepts of decision theory are very hard to put into practice because of the difficulty of evaluating the required probability measures. Therefore, most practical approaches to model-based vision for both specific models, e.g., [Binford, 1982, Bolles and Horaud, 1986, Brooks, 1981, Shneier *et al.*, 1986], and generic models, e.g., [Fischler *et al.*, 1981, Ohta *et al.*, 1979, McKeown and Denlinger, 1984, Huertas and Nevatia, 1988], rely on heuristic measures to select among competing scene parses. These methods, although they may be effective in the context for which they were designed, are extremely hard to extend and require the use of many parameters whose significance is not clearly understood.

On the other hand, approaches such as those of Feldman and Yakimovsky [1974], Georgeff and Wallace [1984], and Rissanen [1983, 1987] provide a sound theoretical basis for the decision problem but offer few practical computational methods for dealing with complex scenes in real images.

In this paper, we focus on an objective function approach to the task of ranking scene-labeling hypotheses. For brevity, we omit discussion of the related problem of hypothesis-generation, and refer the reader to [Fua and Hanson, 1989]. We define a class of objective functions based upon theoretical arguments similar to those of Georgeff, Wallace and Rissanen, and show that the required probability estimates can actually be computed in the context of a few natural assumptions.

Our formulation has many desirable features, but is not by itself a complete solution to the feature extraction problem. To be effective it must be coupled with a robust hypothesis generation mechanism and an efficient optimization procedure. Furthermore, one would like to have models for geometric quality analysis much more complex than those presented here. It should come as no surprise that discovering good *models* and *hypothesis-generation strategies* are the most difficult tasks in the development of a system attempting to perform shape perception. The strength of our approach is that it provides a unified framework that clearly exposes the critical components and characteristics of model-based vision systems.

## 2 Derivation of the Objective Function

The goal of feature extraction is to parse a scene in terms of objects conforming to particular models. To discriminate among competing parses, an objective function must be able to measure the goodness of fit to feature models that include such characteristics as area photometry, edge photometry, shape, and semantic relationships. In this section, we define a basic class of models, discuss the parameters we expect to control our objective functions, derive the theoretical forms of the objective functions themselves, and provide an interpretation of the resulting functions in terms of information theory.

### 2.1 Object Modeling

For the purposes of this work, we define a *model* to be a geometric description of an object in the world characterized by its *geometric constraints* and its *photometric signature*; we define the *evidence* for such objects in digital images to be a collection of *delineated areas* corresponding to major object parts, together with associated quantities directly derivable from the pixel values in such areas.

*This research was supported in part by the Defense Advanced Research Projects Agency under Contract Nos. MDA903-86-C-0084 and DACA76-85-C-0004.

We interpret the photometric signature of any object model in terms of the expected signal from *an ideal object model* plus *a noise model* [Rissanen, 1983, Rissanen, 1987, Leclerc, 1989]. The object's evidence can then be encoded in terms of these models. We will use length of the shortest encoding to measure the quality of the fit between the data and the model.

### 2.2 Essential Parameters of the Objective Function

Our approach introduces two fundamental parameters, the *scale* and the *shape coefficient*:

**Scale.** The scale is interpretable as the unavoidable dimensional factor that converts dimensional quantities such as area or length into dimensionless probabilities. Area units are thus scaled down by two powers of the dimensional unit, while length terms such as edges are scaled down by a single power. The scale parameter thus controls whether the area signature dominates edge signature.

The scale parameter may also be understood by observing that when an image is resampled or zoomed, the area $A$ of a patch will change, but the complexity of the patch, as reflected in its minimal encoding, should remain invariant. Thus there should be some intrinsic zoom factor $s$ that relates the area $A$ to the area $A_0 = A/s^2$ in the zoomed image that has exactly the resolution needed to encode the model complexity without oversampling. The formulas presented later in the paper may thus be alternatively interpreted as expressing the patch encoding cost in terms of the sampling-invariant quantity $A_0$ instead of $A$ itself.

**Shape Coefficient.** An objective function with a shape quality term alone will retain any candidate model instance with the appropriate geometry, even if it does not fit the image data. On the other hand, an objective function with only a photometric model will make the same class of errors as a segmentation algorithm. The shape coefficient balances the possibly conflicting requirements of the geometry and photometry; the point where this balance lies must be determined by the context of the application.

The scale and shape coefficients characterize the fundamental balance of influences that must be semantically specified for each application. Within a particular model domain, it seems possible in principle to estimate the scale by using measures of local complexity. Our approach to feature-hypothesis evaluation provides a clear way to justify and understand the essential role of these two parameters in feature extraction, regardless of the other details of a particular system.

### 2.3 The Probability of a Scene Parse

We choose to describe the problem of determining the best image interpretation as the need to maximize the probability $P = p(m_0 m_1 \ldots m_n | e_1 \ldots e_n)$ that, given the evidence $E = \{e_i;\ i = 1 \ldots n\}$, parsing the scene in terms of a particular set of model instances $M = \{m_i;\ i = 1 \ldots n\}$ and a backround $m_0$ is in fact correct.[1] Each $m_i$ is taken to be a geometric model instance, while $e_i$ is the measurable evidence for the object, typically a collection of associated pixel intensities. Since we are interested in feature extraction, we do not explicitly represent the background and collect no evidence for it.

It is essentially impossible to evaluate the conditional probability $P$ in its most general form, so we make a crucial independence assumption: the probability of a particular model hypothesis is influenced *only* by its corresponding body of evidence and the other model instances. For example, in an aerial image, whether or not a patch of pixels can be identified as a road may depend on its own photometry and on the presence or absence of neighboring houses, but not on the particular photometry of those houses.

Formally, this assumption can be written as follows: If $I, J, K$ denote sets of indices referring to model instances and their corresponding bodies of evidence, we assume $\forall I, J, K$ such that $J \cap I = \emptyset$ and $J \cap K = \emptyset$, $P(m_J e_K | e_I) = P(m_J e_K)$, and $\forall I, J,\quad P(m_J | m_I e_I) = P(m_J | m_I)$.

The assumption may break down when one object's expected photometry is strongly modified by another object, as when a superstructure or a separate building occludes or casts a shadow on a roof. In practice, one can partially compensate for such phenomena by discounting small anomalies.

Combining our assumption with Bayes' rule, it is straightforward to express the probability of the parse as:

$$\begin{aligned} P &= p(m_0 m_1 \ldots m_n | e_1 \ldots e_n) \\ &= p(m_0 m_1 \ldots m_n) \prod_{i=1}^{n} \frac{p(e_i | m_i)}{p(e_i)}. \end{aligned} \tag{1}$$

This expression clearly separates the contribution of the photometry, in the evidence-dependent terms, from the abstract contribution of the geometric and semantic component in $p(m_0 m_1 \ldots m_n)$ under the stated assumption. We further expand this term as:

$$\begin{aligned} p(m_0 m_1 \ldots m_n) &= p(m_0 | m_1 \ldots m_n) p(m_1 \ldots m_n) \\ &= P_0 p(m_1 \ldots m_n), \end{aligned} \tag{2}$$

where $p(m_1 \ldots m_n)$ is the probability that these $n$ models appear in the scene, and $P_0$ is the probability that no other models appear. Since we do not take the background explicitly into account in this work, we consider $P_0$ to be constant.

### 2.4 Minimal Encoding Length and Model Effectiveness

We choose to express the quality of a parse as the (base 2) logarithm [2] of Eq. (1). Classical information theory [Shannon, 1948, Hamming, 1985] leads us to interpret the resulting score $S$ in terms of encoding length:

$$S = +\log \frac{P}{P_0} = F - G, \tag{3}$$

[1] For example, in terms of a human analyst's perception, or in terms of ground truth.

[2] All logarithms in this paper are base 2 logarithms.

where we define

$$F = \sum_{i=1} F_i = \sum_{i=1} \{-\log p(e_i) + \log p(e_i|m_i)\} \quad (4)$$

$$G = -\log p(m_1 \ldots m_n). \quad (5)$$

Here $F$ is what we call the *encoding-effectiveness* of the set of models. The first term in $F$ is the number of bits needed to describe the evidence in the *absence* of the model, while the second term gives the number of bits needed to describe the evidence *in terms of the model.* The term *effectiveness* is thus motivated by the fact that $F$ represents the *number of bits saved* by representing the evidence using the model, and the fact that $F$ increases as the fit improves.

$G$ is the number of bits needed to encode the evidence-free model representation information, and quantifies the elegance of the chosen set of model instances as well as their dependencies.

### 2.5 Remarks

**Feature Extraction Viewed as an Optimization Problem.** The problem of finding the best parse of a scene can now be rephrased as the problem of optimizing over sets of hypotheses evaluated by Eq. (3). Global optimization corresponds to a blind search procedure, which searches all possibilities without attempting to determine which candidates are more likely than others. In practice, the search space may be far too large for this type of search. Since intelligent heuristics can overcome this drawback, a natural way to design an application system is to incorporate hypothesis-generation algorithms that *project* from the space of all possible hypotheses onto a subspace of very likely hypotheses. Such projections have the side effect of reducing the discriminatory burden placed upon the objective function.

**Generic Models Require Photometric/Geometric Balance.** When a model's geometry is completely determined beforehand, as it is for template-matching approaches to automatic shape recognition, there is *no need* for the geometric information component of the objective function, since it is constant and maximum likelihood analysis alone will do. The geometric terms in the objective function begin to play a critical role when we utilize models defined by a set of general geometric constraints in place of a specific shape template. Such *generic models*, with arbitrarily large numbers of parameters, require objective functions like ours that balance their geometric aspects against their photometry.

## 3 Photometry: Computing $F$

Two of the main characteristics of an object in an image are its interior photometry and its contrast with the background, which produces edges. Here we explore simple models for the area and for the edges of an object that have proven useful in analyzing imagery. When working with stereo pairs of images, we also incorporate a stereoscopic model, and compute the depth parameters of an object in the scene by optimizing the corresponding stereo effectiveness.

We have seen that the effectiveness $F$ is computed as $-\log p(e) + \log p(e|m)$ where $e$ represents the grey level values of the pixels that are enclosed by the contour $m$. For the sake of exposition, let us distinguish the evidence $e_A$ relative to the interior of the patch and the evidence $e_E$ relative to the boundary. Formally, we can write:

$$p(e|m) = p(e_A|m)p(e_E|m, e_A)$$
$$p(e) = p(e_A)p(e_E|e_A) \ .$$

We assume that contrast with the background can be measured by using local image derivatives, while ignoring the grey levels of the boundary pixels. This contrast depends on the grey level of background pixels that do not appear in the object descriptions, and can therefore be considered as independent of the interior object photometry. Thus we write $F_i$ in Eq. (4) as the sum of area and edge components:

$$F_i = F_{i,A} + F_{i,E}$$
$$F_{i,A} = -\log p(e_A) + \log p(e_A|m)$$
$$F_{i,E} = -\log p(e_E) + \log p(e_E|m) \ .$$

This prescription must be modified when dealing with objects that share edges, since the contrast of the shared edges is completely determined by the photometry of the regions on both sides of the edge. In this case, the shared boundaries do not contribute to the edge effectiveness term.

When additional images are available and $m$ is a three-dimensional model, additional evidence $e_S$ can be gathered using the projection of $m$ onto each image. We write:

$$p(e, e_S|m) = p(e|m)p(e_S|m, e)$$
$$p(e, e_S) = p(e)p(e_S|e) \ .$$

In the case of a pair of stereo images, $e$ is the evidence measured in the left image and $e_S$ the corresponding evidence in the right image relative to the model projected into that image. For a stereo pair, we therefore add to the effectiveness a *stereo effectiveness* term,

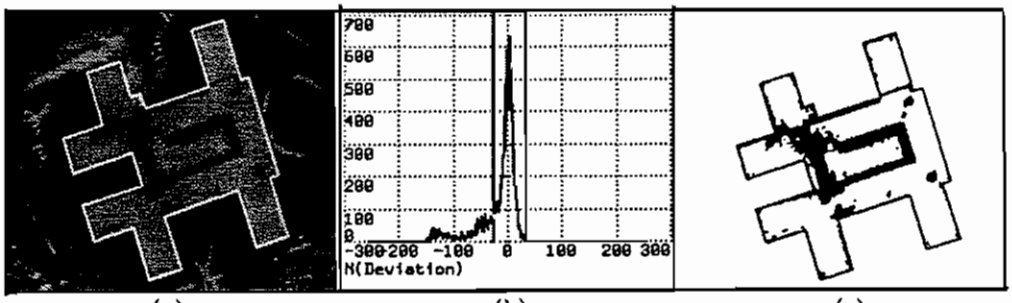
$$F_S = -\log p(e_S|e) + \log p(e_S|m, e) \ . \quad (6)$$

### 3.1 Area Model for Homogeneous Regions

We model the interior intensities of an image region by a smooth intensity surface with a Gaussian distribution of deviations from the surface. Since objects in real images typically have anomalies which do not lie on the smooth surface, we encode such anomalous pixels as outliers. As we shall see later, this can critically enhance the discriminatory power of the area-encoding effectiveness.

In the application of our approach to aerial imagery, we take the intensity surface to be a plane. In Figure 1, we show: (a) An image and a delineated model instance. (b) The histogram of deviations from the planar fit to the intensity surface. (c) The solid white area indicating the location of the pixels within the main Gaussian peak. Black areas within the model outline lie outside the peak and are considered anomalous.

In an 8-bit image, it would take $8A$ bits to encode the pixel values if we did not take advantage of dependencies among pixels. Similarly, it would take $k_A A$ bits




(a) (b) (c).

Figure 1: (a) Delineated model instance. (b) Histogram of deviations. (c) Anomalies.

to encode the same information using our region model, where

$$k_A A = n(\log\sigma + c) + 8\overline{n} + E(n,\overline{n}) \quad . \tag{7}$$

Here $n(\log\sigma+c)$ is the cost of Huffmann-encoding [Hamming, 1985] the pixels in a Gaussian peak, $8\overline{n}$ is the cost of encoding the outliers, and

$$E(n,\overline{n}) = -\left[n\log\frac{n}{A} + \overline{n}\log\frac{\overline{n}}{A}\right]. \tag{8}$$

is the entropy, i.e., the cost of specifying whether a pixel is or is not anomalous. $\sigma$ is the variance of the Gaussian distribution, $n$ is the number of pixels in the Gaussian, $\overline{n} = A - n$, and $c = (1/2)\log(2\pi e)$. Note that in the computation of the encoding cost, we have not included the cost of encoding the six internal parameters of the model: 3 for the plane, 2 for the Gaussian, and one for the probability $n/A$ that a pixel lies in the main peak. It can be shown [Rissanen, 1983, Schwarz, 1978] that these costs are approximately equal to $(1/2)\log A$ bits per internal parameter of the statistical distribution, and are therefore negligibly small compared to $k_A A$.

We weight all areas and lengths using the scale parameter $s$ (see section 2.2) so that the area-encoding effectiveness becomes:

$$\begin{aligned} F_{i,A} &= \text{bits(without model)} - \text{bits(with model)} \\ &= (8 - k_A)\frac{A}{s^2} \\ &= \frac{1}{s^2}\left((8 - c - \log\sigma)n - E(n,\overline{n})\right) \quad . \end{aligned} \tag{9}$$

Optimization of this score is intuitively appropriate because it finds the best compromise among the following:

- large area $A$,
- low standard deviation $\sigma$,
- small number of anomalies $\overline{n}$.

**Effect of Anomaly Discounting.** In the graphs on the left in Figure 2, we plot the area-encoding effectiveness $F_A$ as a function of the radius of a square patch centered at the center of the images shown in the left column: a good but noisy synthetic image of a square, the same image with gross area anomalies, and an image of a similar but distorted square. When we compare the

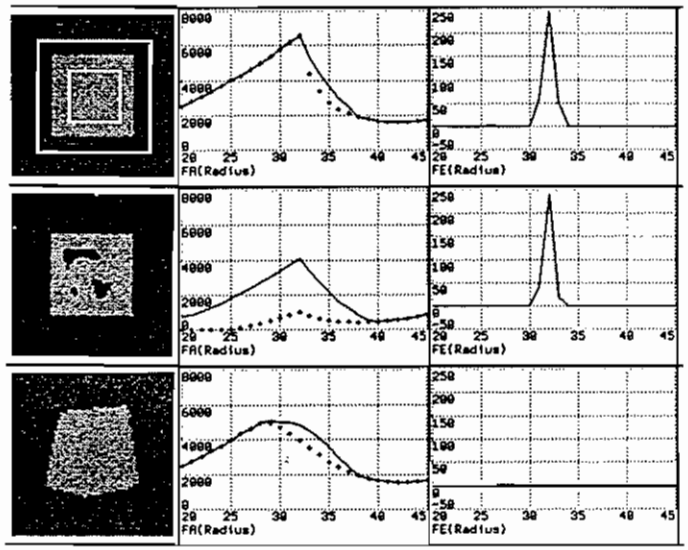



Figure 2: Area and edge effectiveness of a square patch as a function of candidate radius, with (solid) and without (dotted) anomaly discounting.

results obtained *after discounting anomalies* (solid lines) with those results found without anomaly discounting (dotted lines), we see that anomaly discounting *must* be included to make the objective function reliably select the same shape a human observer perceives. This is potentially a critical factor in the practical application of this approach because, as we see in Figure 1, real images nearly always have significant anomalous components.

Note that we only have *local* maxima of the area-encoding effectiveness appearing in Figure 1; for large radii, a better parse of the scene would be in terms of *two* model hypotheses, one square and one square-shaped ring covering the rest of the image, rather than one square plus random background. From this example, we see that high score alone is not an adequate criterion; we must also require local maximality when dealing with a partial description of the scene as opposed to a global one. For this reason it is important in practice to measure whether a candidate object passes this maximality test. Experimentally, we have found that high edge quality enforces this requirement; we now turn to the explicit form of the edge term used.

### 3.2 Edge Model

We adopt the definition [Rosenfeld, 1970, Haralick, 1984, Canny, 1986] of edge pixels as maxima of the local image derivative, and we classify edges according to whether or not an edge boundary pixel conforms to this definition. In the absence of a model, it would take 1 bit per pixel to encode this information. If we now use the 1-parameter model that takes into account the proportion of maximal edge pixels, the most efficient Huffmann [Hamming, 1985] code for this information would require

$$k_E = -\left[\frac{n}{L}\log\frac{n}{L} + \frac{\overline{n}}{L}\log\frac{\overline{n}}{L}\right] \qquad (10)$$

bits per boundary pixel, where $L$ is the length of patch boundary in pixels, $n$ is the number of boundary pixels that are maxima of the local image gradient, and $\overline{n} = L - n$.

We then weight all lengths by the scale factor $s$ and estimate the edge-encoding effectiveness to be

$$\begin{aligned} F_{i,E} &= \text{bits(without model)} - \text{bits(with model)} \\ &= (1 - k_E)\frac{L}{s} \quad . \end{aligned} \qquad (11)$$

As in the case of the area term, we have neglected the $(1/2)\log(L/s)$ bits required to encode the one internal parameter of the model [Rissanen, 1983, Schwarz, 1978].

As shown in the right column of Figure 2, this edge score is maximal when all boundary pixels conform to our edge model, and degrades as the proportion of such pixels diminishes. This model has proven effective in our application of these techniques to aerial images because it provides a measure of edge-quality that does not include an image-dependent threshold on edge strength.

We have also experimented with an edge model that requires the gradient direction be normal to the object outline, and computes the encoding cost of deviations from the normal vector. Both models yield similar rankings.

### 3.3 Stereography

The simplest stereo model assumes that corresponding pixels have the same grey-levels in both images. In practice, to compute the stereo effectiveness of Eq. (6), we determine the number of bits required to encode the projected patch in the second image, while knowing its photometry in the first. We compute the deviations of the intensities from their predicted values and encode them using the same Gaussian model with anomalies that we used for the area term. The anomaly discounting is required because of the possibility of occlusions. We also take into account the edge quality of the contour in the second image and its edge-encoding effectiveness.

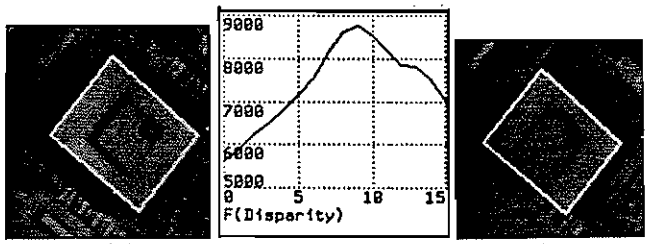
The stereographic effectiveness term $F_S$ is therefore the sum of an edge and an area term:

$$\begin{aligned} F_S &= F_{AS} + F_{ES} \qquad (12) \\ F_{AS} &= (8 - k_{A_2})\frac{A_2}{s^2} \\ F_{ES} &= (1 - k_{E_2})\frac{L_2}{s} \end{aligned}$$

where $A_2$ is the area of the projected patch in the second image, $L_2$ is its boundary length, and $k_{A_2}$ and $k_{E_2}$ are the corresponding model encoding costs.

We can use the effectiveness measure (12) to optimize the elevation parameters of a two-dimensional delineation found in the first image. The search space is extremely constrained since the projected shape is known and the only degree of freedom is epipolar motion in the second image.

Let us consider the stereo pair of images in Figure 3(a,c). Assuming that the roof is horizontal, we plot in Figure 3(b) the value of $F_S$ as a function of the assumed disparity between the candidate outline in the left image (a) and the projected outline in the right image. We note that $F_S$ has a sharp peak for the correct match outlined in (c).




(a) (b) (c)

Figure 3: (a) Roof candidate in left image of a stereo pair. (b) $F_S$ as a function of the assumed disparity between left and right image. (c) The projection of the contour in the right image using the best disparity value.

## 4 Geometry: Computing $G$.

The geometric cost $G$ defined by Eq. (5) is a *measure of quality* of a set of object hypotheses. The simplest way to handle dependencies among objects is to require that there be no conflicts within a particular set of hypotheses; formally we write:

$$\begin{aligned} p(m_i|m_j) &= p(m_i) \quad \text{if } m_i \cap m_j = \emptyset \text{ or } m_i \subseteq m_j \\ &= 0 \quad \text{otherwise} \\ p(m_1 \ldots m_n) &= \prod_i p(m_i) \quad \text{if no conflict} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

It follows that $G$ can be expressed as

$$G = -\log p(m_1 \ldots m_n) = \gamma \sum_{i=1}^{n} G_i, \qquad (13)$$

where $G_i \propto -\log p(m_i)$ is a model quality measure that increases as the shape degrades, and $\gamma$ is the arbitrary *shape coefficient*.

Now we can deduce a mechanism for deciding whether or not the addition of one more feature object is advantageous or detrimental to the overall parse. If we write the overall score in the form

$$F = \sum_{i=1}^{n} (F_i - \gamma G_i),$$

we conclude that we should accept only model instances with $(F_i - \gamma G_i) > 0$, since these are the only ones that improve the likelihood of the full scene parse.

The simplest effective model for $G_i$ is the sum of the cost of chain-encoding the boundary of the object's area plus a constant cost for introducing a new object; this gives a geometric cost

$$G_i = c + \frac{L_i}{s}. \tag{14}$$

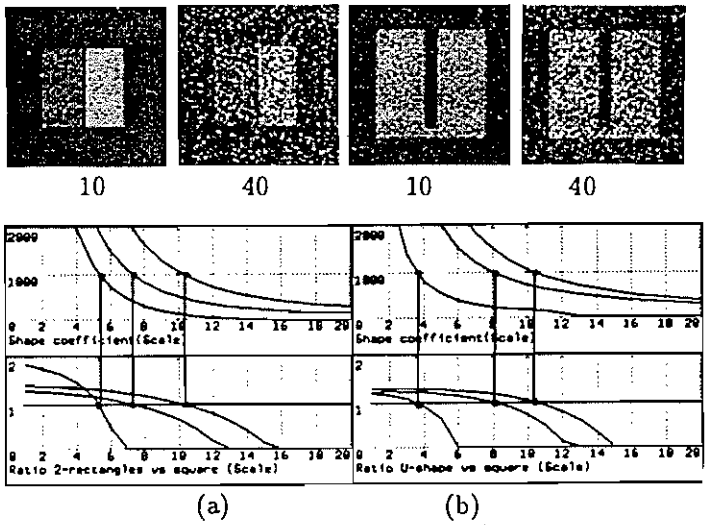



Figure 4: (a) Ratio of single-square to double-rectangle score as a function of noise variance (40, 20, 10). (b) Similar plot comparing the score of the square interpretation to the "U" interpretation.

In Figure 4(a), we show how the length term (14), which gives preference to compact objects, influences the parse when a split square is interpreted alternately as a single compact square or two adjacent rectangles. The bottom graph takes three images, with noise variance 40, 20 and 10, and plots the ratios (two-rectangle score)/(square score) as a function of scale for fixed $\gamma = 1$. Note that increasing the scale in this example amounts to looking at a subsampled image in which fine details are no longer visible. The interesting value of the scale is that for which the scores are *equal*, i.e., the ratio is one. Thus we plot in the upper graphs the locus of points where the ratio is unity as a function of $\gamma$ as well as scale. In Figure 4(b), we carry out a similar plot for an image of a square with a missing portion that makes it "U"-shaped. We see that the ratio ("U" score)/(square score) behaves so that the square interpretation is preferred at a large scale in the best image, and at a much lower scale in the noisier images.

## 5 Examples

We have applied the principle of objective-function optimization to operator-initiated shape extraction and to automated extraction of generic cartographic features such as buildings from aerial imagery, described elsewhere [Fua and Hanson, 1989]. In the automated application, we use an hypothesis generator that carries out the following steps: (1) extract linked edges; (2) find edges obeying geometric constraints (such as rectilinearity) that define enclosed regions in the image; (3) compute the score of each enclosed area using the objective function; (4) find the subset of nonconflicting shape candidates maximizing the total score. One may also optimize each candidate shape with respect to the objective function before the final ranking.

The objective function plays a crucial role in this application because the hypothesis generator will always produce conflicting sets of candidates, and a means of distinguishing among these is absolutely essential.

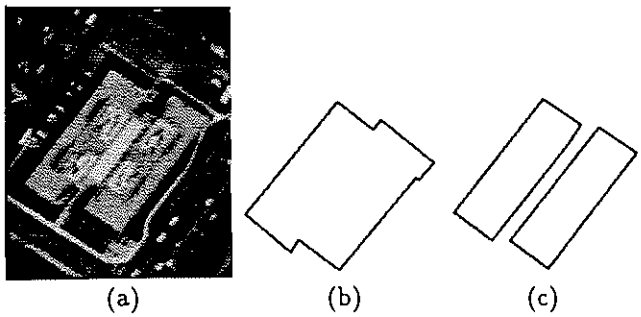



Figure 5: (a) A complex building. (b) Interpretation in terms of a single polygon. (c) Interpretation in terms of two polygons.

For example, for Figure 5(a), the system produces two conflicting interpretations: one in terms of a single polygon enclosing both wings as in Figure 5(b) the other in terms of two polygons, one for each wing as in Figure 5(c). At low scale the latter will be preferred because of its better fit to the photometric data, while at high scale the former will dominate due to its lower geometric cost.

In Figure 6, we show the hypotheses generated and retained by the system for scale values of 6, 7 and 8, with fixed shape coefficient; for this scene, scale 8 clearly gives the best parse.

From the examples shown in this section, we can form an intuitive understanding of the scale parameter: $s$ tunes the scale *not* of the physical size of the object, but the scale of its *quality*. Objects with close fits to the strict model are selected first as we ramp the scale down from a high value.

## 6 Conclusion

In this work, we have shown how an information theoretic approach to the feature extraction problem can be formulated in such a way as to permit realistic computational techniques for the required probability estimates. Our approach provides a firm theoretical basis for understanding complex feature extraction problems that require a balance between photometric evidence and geometric quality. Of course, the objective function approach given here cannot by itself lead to good solutions to the feature extraction problem, but must be teamed

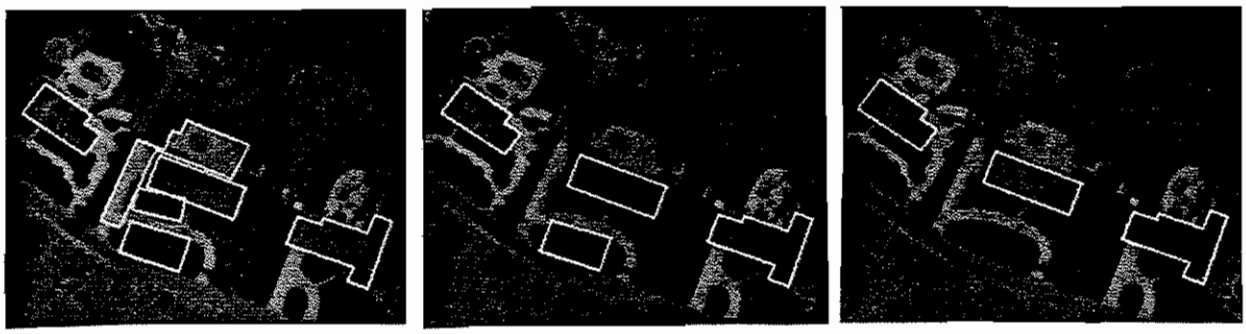


Figure 6: Aerial image of suburban buildings parsed at scales 6, 7, and 8.

with a competent (human or automated) hypothesis generator [Fua and Hanson, 1989]. Among the goals of future work will be the extension of the range of our models and the treatment of complex semantic dependencies in terms of their information-theoretic context.

## References


[Binford, 1982] T.O. Binford. Survey of model-based image analysis systems. *The International Journal of Robotics Research*, 1(1):18–64, Spring 1982.

[Bolles and Horaud, 1986] R.C. Bolles and R. Horaud. 3dpo, a three-dimensional part orientation system. *International Journal of Robotics Research*, 5:3–26, 1986.

[Brooks, 1981] R.A. Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence Journal*, 16, 1981.

[Canny, 1986] J. Canny. A computational approach to edge detection. *IEEE Trans. PAMI*, 8:679–698, 1986.

[Feldman and Yakimovsky, 1974] J.A. Feldman and Y. Yakimovsky. Decision theory and artificial intelligence: I. a semantics-based region analyzer. *Artificial Intelligence*, 5:349–371, 1974.

[Fischler *et al.*, 1981] M.A. Fischler, J.M. Tenenbaum, and H.C. Wolf. Detection of roads and linear structures in low-resolution aerial imagery using a multisource knowledge integration technique. *Computer Graphics and Image Processing*, 15:201–223, 1981.

[Fua and Hanson, 1989] P. Fua and A.J. Hanson. Objective functions for feature discrimination: Application to semiautomated and automated feature extraction. In *Proceedings of the Image Understanding Workshop*, Palo Alto, CA, May 1989.

[Georgeff and Wallace, 1984] M.P. Georgeff and C.S. Wallace. A general selection criterion for inductive inference. In T. O'Shea, editor, *Proceedings of Advances in Artificial Intelligence, Pisa, Italy*, Amsterdam, September 1984. North Holland.

[Hamming, 1985] R.W. Hamming. *Coding and Information Theory*. Prentice Hall, New Jersey, 1985.

[Haralick, 1984] R.M. Haralick. Digital step edges from zero crossings of second directional derivatives. *IEEE Trans. PAMI*, 6:58–68, 1984.

[Huertas and Nevatia, 1988] A. Huertas and R. Nevatia. Detecting buildings in aerial imagery. *Computer Vision, Graphics and Image Processing*, 41:131–152, 1988.

[Leclerc, 1989] Y.G. Leclerc. Image partitioning as constructing simple stable descriptions. *International Journal of Computer Vision*, 1989. in press.

[McKeown and Denlinger, 1984] D.M. McKeown and J.L. Denlinger. Map-guided feature extraction from aerial imagery. In *Proc. 2nd IEEE Workshop on Computer Vision*, pages 205–213, May 1984.

[Ohta *et al.*, 1979] Y. Ohta, T. Kanade, and T. Sakai. A production system for region analysis. In *Proc. 6th Inter. Joint Conf. on Artif. Intell.*, pages 684–686, 1979.

[Rissanen, 1983] J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 2:416–431, 1983.

[Rissanen, 1987] J. Rissanen. Minimum-description-length principle. *Encyclopedia of Statistical Sciences*, 5:523–527, 1987.

[Rosenfeld, 1970] A. Rosenfeld. A nonlinear edge detection technique. In *Proc. IEEE*, volume 58, pages 814–816, 1970.

[Schwarz, 1978] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[Shannon, 1948] C.E. Shannon. A mathematical theory of communication. *Bells Systems Tech J.*, 27:623–656, 1948.

[Shneier *et al.*, 1986] M.O. Shneier, R. Lumia, and E.W. Kent. Model-based strategies for high-level robot vision. *Computer Vision, Graphics and Image Processing*, 33:293–306, 1986.