

# An Integrated Model for Evaluating the Amount of Data Required for Reliable Recognition

Michael Lindenbaum, *Member, IEEE*

**Abstract**—Many recognition procedures rely on the consistency of a subset of data features with a hypothesis as the sufficient evidence to the presence of the corresponding object. We analyze here the performance of such procedures, using a probabilistic model, and provide expressions for the sufficient size of such data subsets, that, if consistent, guarantee the validity of the hypotheses with arbitrary confidence. We focus on 2D objects and the affine transformation class, and provide, for the first time, an integrated model which takes into account the shape of the objects involved, the accuracy of the data collected, the clutter present in the scene, the class of the transformations involved, the accuracy of the localization, and the confidence we would like to have in our hypotheses. Interestingly, it turns out that most of these factors can be quantified cumulatively by one parameter, denoted “effective similarity,” which largely determines the sufficient subset size. The analysis is based on representing the class of instances corresponding to a model object and a group of transformations, as members of a metric space, and quantifying the variation of the instances by a metric cover.

**Index Terms**—Object recognition, localization, pose estimation, similarity measures, noise models, performance analysis.

## 1 INTRODUCTION

MODEL-BASED object recognition and localization are fundamental tasks of computer vision. In localization, one observes a scene, looks for evidence to the presence of a known object in it, and tries to hypothesize its position. In recognition, the object belongs to a known library but is otherwise unknown, and its identity, as well as its position, are to be determined.

Most recognition approaches rely, at least implicitly, on a “hypothesize and test” approach: They hypothesize a particular object and a particular transformation (or “pose”), and check their consistency with the given image. Recognition methods differ mainly in the way they draw the hypotheses, and various methods, such as feature matching, voting in a transformation space, and use of invariants, were suggested. (See the extensive collections [1] and [2].)

A verification process, measuring the consistency between the hypothesized object instance and the given image, directs the decision whether to accept or to reject the hypothesis. The image data usually consists of edge points obtained from brightness discontinuities which, hopefully, correspond to object boundaries. If the hypothesis is correct, many such edge-points should be present near the hypothesized boundary. Note, however, that not all the boundary points are detected due to occlusion, nonoptimal lighting conditions, etc., and the location of the points obtained is usually imprecise, due to noise and imperfect edge detection. Other edge points are also present in the image,

due to other objects and clutter (noise). Therefore, the hypothesis is never consistent with all the data but only with a subset of it.

This raises the natural and fundamental question which is the subject of this paper: What should be required from the consistent data subset, so that the hypothesis could be accepted with high reliability. Intuitively, the larger the consistent set, the more reliable the corresponding hypotheses, but how large is sufficient?

Many reasons may cause the recognition algorithm to fail: The data features in a consistent subset may correspond to another incorrect object or may just be created by the clutter in the image. Even if they correspond to the true object, they may not restrict the pose of the object sufficiently, and the hypothesized object instance may still be associated with some intolerable localization error. All these error mechanisms are addressed by the probabilistic unified framework we propose, which analyzes the different kinds of recognition failures with respect to the combined effect of the object’s shape, the class of transformation allowed, the measurement accuracy, the clutter in the image, and the presence of similar objects in the background. We focus on 2D rigid objects and instances of them obtained using the Affine, Similarity, and Euclidean transformations, analyze the localization and recognition tasks within a common framework, and derive expressions for the size of consistent data subsets, which guarantee the validity of instance hypotheses within arbitrary prespecified confidence.

The amount of information required for reliable verification was considered only in a few papers: Grimson and Huttenlocher analyzed the possibility that a subset of “noise data features” will give false evidence for the presence of an object in the scene [3]. Similar methods are used in the analysis of the reliability of Geometric Hashing

• The author is with the Computer Science Department, Technion, Israel Institute of Technology, Haifa 32000, Israel.  
E-mail: mic@cs.technion.ac.il.

Manuscript received 2 Aug. 1995; revised 4 Sept. 1997. Recommended for acceptance by S. Dunn.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 105664.

technique under various noise models [4], [5]. Maybank derived the distribution of the cross-ratio invariant function, analyzed the use of this invariant as evidence of the object presence, and predicted the “false alarm” rate [6]. Lindenbaum analyzed the effect of the object’s shape on the recognition difficulty and quantified it by two parameters, self-similarity and similarity, that characterize the difficulty of localization and recognition, respectively [7]. More recent papers of Ben-David and Lindenbaum [8], [9] consider this problem from a combinatorial viewpoint, and derive, using learning theory tools, upper bounds on the probability of drawing incorrect hypotheses.

Our work is different from previous contributions in several aspects: First, it is the first approach that considers complicated scenes in which both the effects of object similarities and clutter act to make recognition less reliable. Moreover, here we examine neither the probability that a particular hypothesis, consistent with  $k$  data features, fails, nor the probability that some particular hypothesizing mechanism fails, but, rather, the probability that any hypothesis consistent with  $k$  data features fails. The latter probability, which is naturally higher, sets an absolute bound on the error irrespective of the particular algorithm and, in particular, of the way it chooses its hypotheses. In this sense, our approach is algorithm independent. Furthermore, the objects are characterized only by an upper bound on their perimeter, and are neither required to be polygonal [3] nor assumed to be represented only by a collection of feature points [4], [5]. The derivation is rigorous and no assumptions, except the modeling of the data itself, are made. This characteristic implies that all the results are guaranteed, but requires some price in the form of looser bounds.

We start by presenting a probabilistic model for the available data features, which incorporates the presence of objects, as well as the inevitable appearance of clutter. The analysis starts by considering the simplified scene considered in [3], which consists only of data features originated by clutter. Then, we assume that no clutter is present in the image, and examine recognition failures due to the presence of objects in the scene. Finally, the two effects, of clutter and presence of objects, are combined into one integrated model for predicting recognition failures, yielding expressions for the sufficient number of consistent features which guarantee a reliable hypothesis, both for localization and for recognition.<sup>1</sup> Then, we derive related figures, including a bound on the data subset size, required for preventing object misses and draw the resulting ROC curve. Finally, we present simulation results, which supports the analytical derivation.

## 2 MODELING THE DATA AND THE RECOGNITION PROCESS

### 2.1 The Data Collection Model

We consider recognition processes preceded by an edge-detection preprocessing stage, which provides the location

of boundary points. Commonly, such edge-detection processes are unreliable and provide inexact locations. This imperfection is modeled as follows: Every data feature (edge point) originates either by the boundaries of objects that are present in the scene, or by the general clutter (noise) in it. Data features, originated by the image clutter, are randomly located in the image according to a uniform density. Data features, originated by the objects’ boundaries, are associated (according to our model) with a bounded error, and their generation model is slightly more complicated: Let  $\partial V_t$  be the boundary of an instance of the object  $V$ , corresponding to a transformation  $t$ . We assume that data points extracted from this boundary are independently sampled according to a uniform distribution, inside

$$V_t^\Delta = \{r \mid \exists s \in \partial V_t \text{ s. t. } \|s - r\| < \Delta\}, \quad (1)$$

which we refer to as “extended boundary.” (This planar set may be also described as the morphological dilation of  $\partial V_t$  with a circular structuring element.)

We assume that the data features set,  $S$ , is originated by both sources, and that data features are points that are randomly drawn according to the following piecewise uniform distribution: In regions that belong to some extended boundary of some object in the image, the distribution density,  $f_b$ , is constant and higher than the clutter density,  $f_c$ , in the rest of the image.  $f_c < f_b$ . (The subscript “b” denotes “boundary,” and the subscript “c” denotes “clutter.”) We further assume that the objects observed in the image are in relatively random locations, and denote the fraction of area occupied by their extended boundaries by  $\alpha$ . For an  $l \times l$  image, the constraint  $\int [\alpha \cdot f_b + (1 - \alpha) \cdot f_c] = 1$ , obtained by integrating the density over the image, readily follows.

The simple model, suggested above, does not cover all situations in computer vision. It addresses, however, important issues, such as the dependence of the data on the objects in the scene, the uncertainty on the observed part of the object, and the inaccuracy of the measurements. We argue that the images produced by the model have much in common with real edge images, although they do not look exactly like them (see Fig. 1 for a synthetically generated example of such data). To our best knowledge, this is the most general model considered, so far, for general edge images, which lends itself to rigorous analysis without approximations.

Our analysis assumes only that the distributions are bounded and can be easily extended to nonuniform sampling densities (see, e.g., the treatment of occlusion in [7]). The treatment of other particular distributions (e.g., Gaussian, see [5]) was not considered but should be similar.

### 2.2 Some Assumptions on the Recognition Process

The hypotheses drawn by the recognition algorithm depend on the observed data, the objects believed to be present in the scene, and the transformations allowed, which reflects our knowledge on the imaging system. Here, we focus on 2D rigid objects and on affine transformations. We refrain from referring to any particular method for drawing the hypothesis about the object identity and pose, and assume that any object instance that is consistent with subsets of the data of size  $k$  or more may be hypothesized. That is,

1. Due to the lack of space, only outlines of the proofs are given here. For the full proofs, see the full version [10].

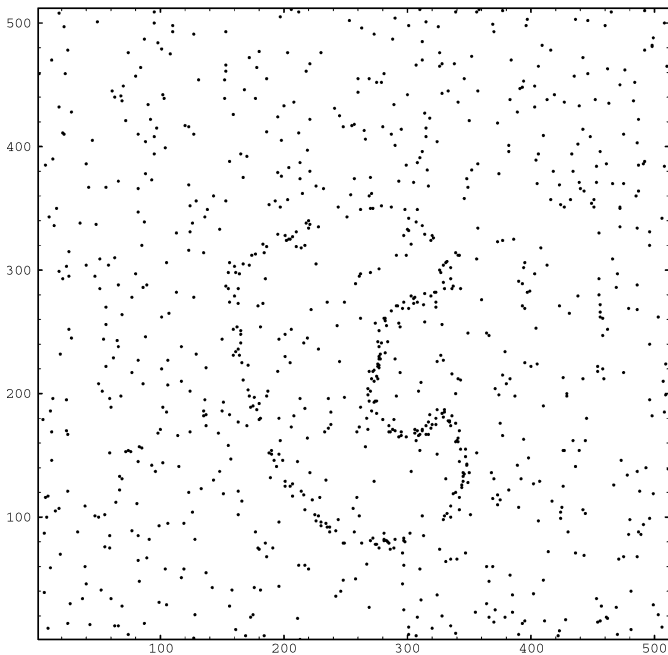


Fig. 1. A synthetic collection of data features created according to the *mixed data model* (see definition later), assuming one object is present in the scene (parameters  $s = 5$ ,  $\alpha = 0.085$ ). Note the accumulation of points near the boundaries of the object in the image center.

the hypotheses  $V_t$ , which corresponds to the object  $V$  and the transformation  $t$  may be drawn by the recognition algorithm if

$$\|S \cap V_t^A\| \geq k. \quad (2)$$

Such an algorithm may fail, however, if some sufficiently large subset of data features accidentally matches some instances of an object, leading to a hypothesis which is incorrect either in object pose or identity. The probability for such an event decreases if the required size  $k$  of the consistent subset increases. Note, however, that although this approach is common and used in almost all existing verification procedures, it may not be the optimal one. See [5] for an approach that builds on a Gaussian distribution, and [11], [12], where taking into account occlusion model and grouping information was shown to improve the false alarm rate.

Practical recognition algorithms draw their candidate hypotheses out of some discrete version of the hypothesis space, usually obtained either by explicit discretization of the transformation parameters (e.g., Hough transform) or some data-driven implicit discretization (e.g., Geometric Hashing). Here, we do not consider any such discretization, but assume that every hypothesis that satisfies (2) is accepted. Therefore, in our analysis of false hypotheses, we consider all hypotheses made by any particular algorithm, and many more, which makes our results more conservative than those derived in the context of particular algorithms.

### 3 ANALYSIS OF THE OBJECT LOCALIZATION TASK

#### 3.1 The Goal and the Main Result

This section analyses localization or pose-estimation procedures. Its main result is an expression for the sufficient size

of a data subset, that, if consistent with an instance of the object  $V$ , guarantees<sup>2</sup> that this instance is a sufficiently good hypothesis of the correct object instance. By the title “sufficiently good hypothesis,” we mean that it is an instance of the correct object, and that its pose is close enough (in a quantitative sense defined later) to the pose of the correct instance. Naturally, the expression depends on image quality factors, such as the amount of clutter in the image and the saliency (or visibility) of the object’s boundary, quantified by the  $\alpha$ ,  $f_b$ , and  $f_c$  parameters of the above image model. It also depends on the similarity between  $V$  and other objects that may appear in the scene, on the required localization precision, on the required confidence in this decision, and on a specific type of object symmetry, associated with  $V$ .

Note that this sufficient size is different from the size of consistent data set sufficient to guarantee that a *particular hypothesis* is true. Knowing the latter number is useful for deciding whether to accept a particular hypothesis during the execution of some particular algorithm (see Sarachik [13]). The data set size, which we look for, guarantees a stronger implication: That *every hypothesis* consistent with this data size is successful (in a well-defined sense). This implication is stronger because the probability that one hypothesis out of many is false is higher than the probability that a particular hypothesis fails.

Therefore, any algorithm which relies on accepting hypotheses consistent with this sufficient subset size is bound not to make any false hypotheses, irrespective of the way it chooses its hypotheses. Most algorithms differ only in the way they produce the hypotheses, and not in the way they evaluate them, implying that this analysis apply for them.

EXAMPLE 1. To illustrate the exact meaning of the sought-for sufficient subset size, consider the particular case of a “continuous Hough transform.” Let  $L(\rho, \theta)$  denote the line  $\{(x, y) \mid x \cos \theta + y \sin \theta = \rho\}$ . Consider an image containing  $N$  uniformly distributed random data features, associated with an isotropic measurement error bounded by  $\Delta$ . Then, every point  $(x_i, y_i)$  in the image corresponds to a parameter region

$$K_{(x_i, y_i)} = \{(\rho, \theta) \mid \exists (x, y) \text{ s.t. } (x, y) \in L(\rho, \theta) \text{ AND } \|(x, y) - (x_i, y_i)\| \leq \Delta\},$$

which contains all parameter pair corresponding to lines passing  $\Delta$ -close to that point. The Hough transform operation is essentially to find the maximum of the function  $H(\rho, \theta) = \sum_{i=1}^N K_{(x_i, y_i)}$  on a regular grid in the  $(\rho, \theta)$  space. The scene associated with this example contains no line, implying that this maximum should not be too high. The sufficient subset size provided here bounds the maximal value of  $H(\rho, \theta)$ , which does not necessarily correspond to any  $(\rho, \theta)$  value included in a grid, but clearly bounds the

2. In our probabilistic framework, whenever we say “guarantee,” we mean “guarantee with some prespecified confidence.”

maximal value obtained on any grid (the case of a particular algorithm: the common, discrete, Hough transform), as well as the value of  $H(\rho, \theta)$  for any particular  $(\rho, \theta)$  pair (the case of a particular hypothesis).

To make the derivation clearer, we start with a simpler scenario, in which the image contains only clutter, that is, uniformly distributed random data features. We find the sufficient consistent subset size for this image model. Then, we treat the opposite type of scenario, in which the images contain only data features from the boundary of an object, and show the effect of the shape related parameters, namely, the similarity between the true object and the other objects, and the symmetries of the objects (quantified by a “self-similarity” parameter). Finally, we show how to combine the effects of clutter and similarities to find the sufficient size of the consistent data set for the general, combined, image model.

The following claim, summarizing the main result of this section, contains some unfamiliar parameters, which are discussed and clarified latter.

**THEOREM 1.** *Let  $S$  be a set of  $N$  random features drawn in an  $l \times l$  image according to the mixed data model. Consider a localization procedure which relies on consistent data subsets of size  $k$  and tries to localize the object  $V$  under the affine transformation group  $\mathcal{A}$  with required localization precision  $d_0$ . Let  $r_{total}$  be the total similarity parameter that characterizes the task. Then, the localization procedure is guaranteed, with confidence  $1 - \delta$ , not to produce any false hypothesis, if*

$$k \geq \frac{1}{\log_e(v+1)} \left[ vNp(s, \alpha, r_{total}) + 6 \log_e \frac{evNs}{6(1+\alpha(s-1))} + \log_e c(\mathcal{A}, V, \Delta) + \log_e \frac{1}{\delta} \right], \quad (3)$$

where  $c(\mathcal{A}, V, \Delta)$  is given in (10),  $v$  is any positive number,  $s = f_b/f_c$ , and

$$p(s, \alpha, r_{total}) = \frac{1}{l^2} \max_{t \in \mathcal{A}} \text{area}(V_t^\Delta) \left[ 1 + \frac{r_{total}(s-1)(1-\alpha)}{1+\alpha(s-1)} \right].$$

The consistent set size depends on the similarities between the hypothesized object  $V$  and the objects in the scene (and also on the similarity of the object  $V$  to itself, in a sense to be described), via the total similarity parameter,  $r_{total}$ , which quantifies the effect of similarities in a meaningful way. The claim actually provides a sequence of valid bounds, depending on the parameter  $v$ , which can take any positive value. (See Fig. 2 for a plot of these bounds for a particular scene described later in Example 2.) Therefore, the parameter  $v$  should be optimized to provide the tightest (lowest) bound.

The rest of this section is dedicated to deriving this expression.

### 3.2 Recognition Failures Due to Image Clutter

Consider a simplified scenario, containing only  $N$  data features that are drawn according to a uniform distribution (similar to [3]). This image contains no object, and, therefore, any consistency of the image features with some particular instance of a particular object is accidental. We

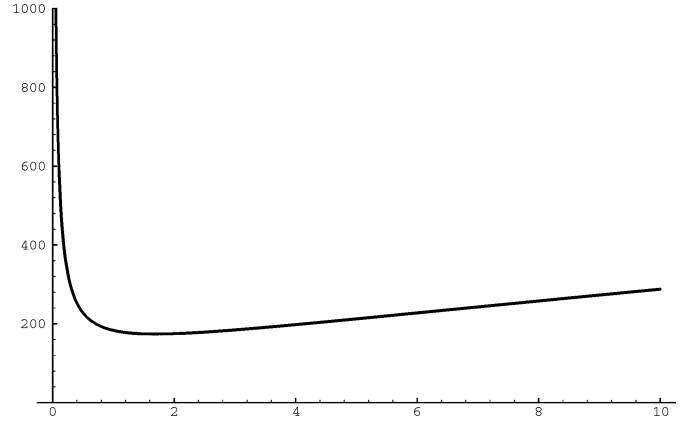


Fig. 2. This figure illustrates the fact that the claims provide many bounds. It corresponds to the clutter only scenario described in Example 2. Every vertical coordinate of a point on the curve is a valid bound corresponding to the value of the parameter  $v$  in the corresponding horizontal coordinate. The tightest (lowest) bound can be chosen from the graph and, here, it is 174 (corresponding to  $v = 1.69$ ).

would like to know what the probability is that the data set  $S = (s_1, s_2, \dots, s_N)$  contains subsets (of certain size) that are consistent with some (unknown) instance of a known object (the event is referred to as localization false alarm because only one object model is considered). Let the class of extended boundaries (instances) associated with a certain object  $V$  and a group of transformations  $T$  be

$$C_{T^\Delta}(V) = \{V_t^\Delta \mid t \in T\} \quad (C_T(V) = \{V_t \mid t \in T\}). \quad (4)$$

Our goal is to find the probability that some subset of  $k$  data features is included in some extended boundary which belongs to  $C_{T^\Delta}(V)$ ,

$$\text{Prob}\{FA_{clutter}\} = \text{Prob}\{\exists V_t^\Delta \in C_{T^\Delta}(V) \text{ s.t. } \|S \cap V_t^\Delta\| \geq k\}. \quad (5)$$

Having an expression for this probability, we can set a value to  $k$  so that such an accidental consistency happens with any required low probability  $\delta$ . We start with an even simpler question, related to a *particular* object instance.

#### 3.2.1 False Alarm Due to a Particular False Instance

Let  $V_i$  be a particular instance of the object. Then,

$$\text{Prob}\{s_i \in V_i^\Delta\} = \frac{\text{area}(V_i^\Delta \cap I)}{\text{area}(I)} \leq \frac{2L\Delta}{l^2} = p_0,$$

where  $I$  denotes the  $l \times l$  image, and  $L$  is an upper bound on the perimeter of any instance of  $V$ . Now, the Hoeffding bound [14] implies that

$$\text{Prob}\{\|S \cap V_i^\Delta\| \geq k\} \leq \left(\frac{k_c}{k}\right)^k e^{-k_c}, \quad (6)$$

where  $k_c = N \cdot p_0$  is an upper bound on the expected number of features in  $V_i^\Delta$ ,  $E\{\|S \cap V_i^\Delta\|\}$ . (The subscript  $c$  in  $k_c$  stands for “clutter.”) The expression (6) matches our intuition well: It increases with the size of  $S$  and the area of  $V_i^\Delta$ , and becomes lower when  $k$  is increased.

### 3.2.2 Representing the Instances as Members of a Metric Space

Calculating (5) from (6) could have been easy, if the class  $C_{T^\Delta}(V)$  was finite. Then,

$$\begin{aligned} \text{Prob}\{FA_{\text{clutter}}\} &= \text{Prob}\{\exists V_t^\Delta \in C_{T^\Delta}(V) \text{ s.t. } \|S \cap V_t^\Delta\| \geq k\} \\ &\leq \|C_{T^\Delta}(V)\| \max_{t \in T} \text{Prob}\{\|S \cap V_t^\Delta\| \geq k\}. \end{aligned} \quad (7)$$

This technique fails here because the group of affine transformations, and the corresponding instance class, is continuous and infinite. Treating all these instances and their corresponding events independently leads to an infinitely high “bound.” Our intuition, however, tells us that not all instances in the class  $C_{T^\Delta}(V)$  are very different from each other, and that the independence assumption is wrong for, say, two extended boundaries of similar instances, which occupy approximately the same region in the plane. This intuition may be quantified and used by treating the (continuous) set of instances as members of a metric space and by constructing its (discrete) cover.

A metric space  $(X, d)$  is a set  $X$  associated with a distance function  $D(\cdot, \cdot)$  between its members, satisfying, for every  $x, y, z \in X$ , three conditions:

- 1)  $D(x, x) = 0$ ,
- 2)  $D(x, y) = D(y, x) \geq 0$ ,
- 3)  $D(x, z) \leq D(x, y) + D(y, z)$ .

(We use a weaker definition of a metric space, which is often referred to also as “pseudo metric.”)

Here, we shall use a distance measure between extended boundaries, which is induced by the random feature distribution. The distance between two extended boundaries (or other planar subsets) is defined as the probability of a (random) feature to fall in their symmetric difference.

$$D(V_t^\Delta, V_{t'}^\Delta) = \text{Prob}\{s_i \in V_t^\Delta \Delta V_{t'}^\Delta\}. \quad (8)$$

For the uniform density considered here, this distance is proportional to the symmetric difference area. The triangle inequality, as well as the other required properties, holds. Therefore, this distance measure is a metric and the set of extended boundaries is a metric space.<sup>3</sup>

A  $\eta$ -cover of a subspace  $A \subset X$  of the metric space is another subset  $A' \subset X$ , such that, for every member of  $A$ , there is at least one member of the cover  $A'$  which is  $\eta$ -close. Note that the cover may be discrete and finite even if the approximated subset,  $A$ , is not. A particular type of cover, specified below, is well defined if there is an inclusion relation  $x \subset y$  between the members of the metric space. Here, these members are subsets of the plane, and the inclusion relation clearly exist.

**DEFINITION 1** (enclosing cover). *Let  $X$  be a metric space, the members of which are subsets of some other space, and let  $D(\cdot, \cdot)$  be the metric associated with it. An enclosing cover of a subspace  $A \subset X$  is another subspace  $A' \subset X$  such that,*

3. A symmetric difference between two sets,  $A$  and  $B$ , is the set  $(A \cap \bar{B}) \cup (B \cap \bar{A})$ . The symbol  $\Delta$  is used to denote both the symmetric difference operator, and, when used as a superscript, the extended boundary. The meaning should be evident from the context.

for every member  $x \in A$ , there is at least one member  $y \in A'$  which satisfies

- 1)  $D(x, y) \leq \eta$ ,
- 2)  $x \subset y$ .

The next stage is to specify a cover for the extended boundaries set. The cover depends on the Transformation class, and is specified for the affine transformation class  $T = \mathcal{A}$ , by the following lemma.

**LEMMA 2.** *Consider the metric space of planar subsets, associated with a symmetric difference distance, which is induced by bounded distribution density  $f_0(\cdot) \leq f_0^+$ . Then, the set of extended boundaries,  $C_{\mathcal{A}^\Delta}(V)$ , has an  $\eta$ -enclosing cover,  $C'_{\mathcal{A}^\Delta}(V)$ , of size*

$$\|C'_{\mathcal{A}^\Delta}(V)\| \leq c(\mathcal{A}, V, \Delta) \cdot (I^2 f_0^+)^6 \cdot \frac{1}{\eta^6}, \quad (9)$$

where

$$c(\mathcal{A}, V, \Delta) = \frac{1}{2} \cdot 6^6 (1 + L^\Delta/2)^2 (L^\Delta)^6 L^4 T^{-12}, \quad (10)$$

and  $L^\Delta = L + 2\pi\Delta$ .

**PROOF** (sketch). The cover members are “extended boundaries” with width higher than  $\Delta$ . Every one of them contains an infinite number of (“regular”) extended boundaries associated with  $\Delta$  width. The cover members are associated with transformation parameters selected on a grid in the 6D transformation space. The grid is fine enough so that every member of  $C_{\mathcal{A}^\Delta}(V)$  is included in some cover member. A larger width of the extended boundary associated with the cover member makes the cover less accurate (higher  $\eta$ ) but smaller.  $\square$

The cover size grows with its precision  $\eta$ , the “variance” of the distribution (the  $(I^2 f_0^+)^6$  term), and the constant  $c(\mathcal{A}, V, \Delta)$ , which is independent of both of them. For the uniform distribution, the size of the cover is just  $c(\mathcal{A}, V, \Delta)\eta^{-6}$ . (The more general form is useful later.) Note also that the cover size can be roughly expressed as a sixth power of some function. For less general transformations (Translations, Euclidean, and Similarity), the cover size is a lower power of similar expression (second, third, and fourth power, respectively). Thus, the cover size quantitatively conveys the effect of degree of freedom allowed by the transformations.

### 3.2.3 Consistency with an Arbitrary Instance of an Object

The use of the enclosing cover allows us to analyze the continuous class of instances by the properties of its finite cover, and to derive the sufficient subset size. The lemma is phrased more generally than required for analyzing the cluttered image case, so that it can be used later for other cases.

**LEMMA 3** (main). *Let  $f_0(\cdot) \leq f_0^+$  be a bounded distribution density. Let  $S$  be a set of  $N$  random features drawn according*

to  $f_0(\cdot)$  within an  $l \times l$  image. Let  $C_{T^\Delta}(V)$  be the metric space of extended boundaries, associated with a symmetric difference metric weighted by  $f_0$ , and let  $C'_{T^\Delta}(V)$  be a  $\eta$ -enclosing cover of this space ( $0 \leq \eta < 1 - p_0$ ). Let  $p_0 = \max_{t \in T} \text{Prob}\{s_t \in V_t^\Delta\}$ . Then, the probability that some subset of  $k$  data features is consistent with some extended boundary in  $C_{\mathcal{A}^\Delta}(V)$  is smaller than  $\delta$  provided that

$$k \geq \frac{1}{\log_e(v+1)} \left[ vN(p_0 + \eta) + \log_e \|C'_{T^\Delta}(V)\| + \log_e \frac{1}{\delta} \right]$$

(where  $e$  is the natural logarithm base and  $v$  is any positive constant).

**PROOF (sketch).** The probability that a large number of data features fall in some extended boundary is smaller than the probability that these features fall in the cover member which includes it. Therefore, the probability that no object instance is consistent with  $k$  features is larger than the probability that no member of the cover includes  $k$  features. The latter may be found by (7), as the cover is finite.  $\square$

The bound depends on the cover size and on the additional parameter  $v$ . Therefore, the lemma actually provides many bounds, one for every pair of the parameters  $\eta$ ,  $v$  in the specified range ( $0 \leq \eta < 1 - p_0$ ,  $v > 1$ ), and they are all valid. Note also that the bounds qualitatively agree with intuition: They grow with the density of data (which increases with  $Np_0$ ), with a demand for increasing reliability (smaller  $\delta$ ) and with the number of "possible alternatives." (The actual number of alternatives, or object instances, is infinite, but, as mentioned above, the cover size depends on the class of transformations and indicates that "more" alternatives are associated with the more general transformation classes.)

We now return to the case of clutter-only image, where the data features are drawn by a uniform random distribution, and optimize over  $\eta$  (for every value of  $v$ ).

$$\begin{aligned} \frac{d}{d\eta} \left[ \frac{1}{\log_e(v+1)} \left[ vN(p_0 + \eta) + \log_e c(\mathcal{A}, V, \Delta) \eta^{-6} + \log_e \frac{1}{\delta} \right] \right] \\ = \frac{1}{\log_e(v+1)} \left[ vN - \frac{6}{\eta} \right] = 0 \Rightarrow \eta_{opt} = \frac{6}{vN}. \end{aligned} \quad (11)$$

Inserting the optimal  $\eta$  value, we get the following claim:

**CLAIM 4.** Let  $S$  be a set of  $N$  random features drawn according to a uniform distribution within an  $l \times l$  image. Let  $C_{\mathcal{A}^\Delta}(V)$  be a class of extended boundaries that correspond to the object  $V$  and the affine transformation class. Then, the probability that some subset of  $k$  data features is consistent with some extended boundary in  $C_{\mathcal{A}^\Delta}(V)$  is smaller than  $\delta$ , provided that

$$k \geq \frac{1}{\log_e(v+1)} \left[ vNp_0 + 6 \log_e \frac{evN}{6} + \log_e c(\mathcal{A}, V, \Delta) + \log_e \frac{1}{\delta} \right], \quad (12)$$

where  $p_0$ , the (maximal) probability that a particular data

feature is consistent with a particular instance of the object  $V$  is  $\frac{2(L+2\pi\Delta)}{l^2}$ ,  $e$  is the natural logarithm base, and  $v$  is any positive constant.

**EXAMPLE 2.** Consider a  $512 \times 512$  image ( $l = 512$ ) containing  $N = 10,000$  random point data features. We would like to know what is the number  $k$  of data features that guarantee, with confidence of 0.999 ( $\delta = 0.001$ ), that no affine instance of a given object  $V$  is consistent with  $k$  or more of the data features. Assuming an isotropic measurement error bounded by  $\Delta = 2$  and that the perimeter of any affine instance is smaller than 400 (a natural assumption which limits the magnification), we may use (12) to calculate the bound as a function of the parameter  $v$  and choose the  $v$  value which yields the lowest bound, which is 174 (see Fig. 2).

### 3.3 Recognition Failures Due to Object Similarities

Besides clutter, real images contain objects which cause the data distribution to be nonuniform and object dependent. In the context of testing a particular hypothesis relative to a given image, it may be sufficient to estimate the feature density locally (see Sarachik [13]). Here, we take a different approach, model the effect of other objects on the feature density explicitly, and show how it influences the probability of false hypothesis.

The approach we take is similar to the previously described analysis of false alarms due to clutter. The main difference is that, here, the data features are placed according to a distribution which depends on the real object instances in the scene, and both the probability

$$\text{Prob}\{x \in V_t^\Delta \mid \text{some data model}\}$$

and the metric cover are modified accordingly.

First, consider a clutterless scene containing only a single object instance  $W_t$  of the object  $W \neq V$ . In this case, all the  $N$  feature points are drawn according to a uniform probability density  $f_b$  within the extended boundary  $W_t^\Delta$ . The localization procedure may make a *discrimination error* if a large data subset is consistent with some instance of the object  $V$ . The probability of such an error largely depends on the probability that a feature point is included in some particular extended boundary of the object  $V$ , and that probability is bounded as follows:

$$\begin{aligned} \text{Prob}\{x \in V_t^\Delta \mid W_t\} &= f_b \cdot \text{area}(V_t^\Delta \cap W_t^\Delta) \\ &= \frac{\text{area}(V_t^\Delta \cap W_t^\Delta)}{\text{area}(W_t^\Delta)} \\ &\leq \max_{t \in T} \frac{\text{area}(V_t^\Delta \cap W_t^\Delta)}{\text{area}(W_t^\Delta)}. \end{aligned} \quad (13)$$

The area ratio  $\max_{t \in T} \frac{\text{area}(V_t^\Delta \cap W_t^\Delta)}{\text{area}(W_t^\Delta)}$  is denoted  $r_{\max}^{W,V}(\Delta)$ , and is named *Similarity* [7]. It quantifies the similarity of the objects regarding their shape, the allowed transformations,

and the observation precision  $\Delta$ . A low, close-to-zero, value of the similarity parameter indicates that the instances of  $V$  and  $W$  are always very different visually, and that a small number of data features probably suffice to discriminate between them. A high, close-to-one, value indicates, on the other hand, that there are instances of  $V$  and  $W$  which are very similar, which makes the discrimination between them difficult.

Note that the *Similarity* parameter depends on the particular instance  $W^{t'}$  in the scene. It would be more elegant if we could define the *Similarity* independently of the instance that is, by  $\max_{t,t' \in T} \frac{\text{area}(V_t^\Delta \cap W_{t'}^\Delta)}{\text{area}(W_{t'}^\Delta)}$ . This, however, is impossible in the general affine case, because there are always transformations  $t$  and  $t'$  that take both  $W$  and  $V$  to the same straight line segment, implying that the similarity is one! This limitation is not a fault of our scheme but is inherent to the recognition task—it is simply impossible to recognize planar shapes from a viewing direction parallel to the shape plane. This problematic issue was also raised by Rucklidge [15], where the algorithm is simply ignoring the problematic transformation subspaces. Note that, for Euclidean and Translation transformations, the similarity is independent of  $t'$ .

Similarly, we may consider a situation where it is an instance  $V_t$  of the hypothesized object  $V$  itself, which is present in the scene. The localization procedure may still fail by hypothesizing some instance of  $V$  which is associated with an intolerable localization error. Such a *localization error* is made if a large data subset is consistent with such an instance of  $V$ . To treat the difficulty of the localization task in a meaningful and quantitative way, one should specify the *localization precision* required. A general and quantitative way to do that would be to choose some distance function  $d(\cdot, \cdot)$  between two instances, and some threshold value  $d_0$ . Then, for every hypotheses  $V_t$  that is considered to be close enough,  $d(V_t, V_{t'}) \leq d_0$ . The choice of the distance measure depends on the user preconception of the nature of successful localization, and also depends on the application. In [7], where this quantification of localization performance was suggested, it was argued that a specific metric between instances, the maximal distance between corresponding boundary points, may be preferable for robotics applications. We emphasize here that any other distance measure (e.g., Hausdorff metric, difference of area metric, etc.) may be chosen as well, and do not consider the choice of the distance metric further.

Localization fails if instance  $V_t$  associated with a large localization error is hypothesized. The class of the extended boundaries associated with these “far” object instances,

$$C_{T_{(d>d_0)}^\Delta}(V) = \{V_t^\Delta \mid d(V_t, V_{t'}) > d_0; t \in T\} \quad (14)$$

naturally depends on the particular instance  $V_{t'}$ .

The probability of such a *localization error* largely depends on the probability that a feature point is consistent with some particular “far” instance of the object  $V$ . This probability is bounded as follows:

$$\begin{aligned} \text{Prob}\left\{x \in V_t^\Delta \mid V_t^\Delta, V_{t'}^\Delta \in C_{T_{(d>d_0)}^\Delta}(V), V_{t'}^\Delta\right\} &= f_b \cdot \text{area}(V_t^\Delta \cap V_{t'}^\Delta) \\ &= \frac{\text{area}(V_t^\Delta \cap V_{t'}^\Delta)}{\text{area}(V_{t'}^\Delta)} \\ &\leq \max_{t \in T} \frac{\text{area}(V_t^\Delta \cap V_{t'}^\Delta)}{\text{area}(V_{t'}^\Delta)} \\ &= r_{\max}^V(\Delta, d_0). \end{aligned} \quad (15)$$

The parameter  $r_{\max}^V(\Delta, d_0)$ , introduced in [7], is denoted *Self-Similarity*. It quantifies the minimal difference in the appearance of instances that are far from each other, and depends on the object’s shape, the observation precision  $\Delta$ , and the class  $T_{(d>d_0)}$ , of transformation allowed (and, therefore, on both  $d(\cdot, \cdot)$  and  $d_0$ ). It is expected that instance pairs which are far from each other are not similar, in the sense that the symmetric difference between them has large area. This observation is generally true but fails for many “close to symmetric” objects, implying that discrimination between their instances is hard. Consider, for example, a circular object with a small wedge on its edge. The self-similarity of such an object is close to one even for transformations that rotate the object by  $180^\circ$  degrees, which corresponds to a large distance according to many metrics. This just reflects the observation that many measurements are required for localizing such an object. (See [7] for a discussion of the similarity and the self-similarity properties.)

Now, we may use the *similarity* and the *self-similarity* parameters to find the sufficient data set size required for localization. The following corollary, concerned with discrimination between different object, is proved by a direct application of Lemmas 2 and 3. Similar results concerned with preventing inexact pose hypotheses are easily obtained by replacing the similarity parameter,  $r_{\max}^{W,V}(\Delta)$ , in Corollary 5 by the self-similarity parameter,  $r_{\max}^V(\Delta, d_0)$ .

**COROLLARY 5.** *Let  $S$  be a set of  $N$  random features drawn according to a uniform density  $f_b \approx \frac{1}{2\Delta}$  inside the extended boundary of some object instance  $W_t$ . Let  $C_{\mathcal{A}^\Delta}(V)$  be a class of extended boundaries that correspond to the object  $V$  and the affine transformation class. Then, the probability that some subset of  $k$  data features is consistent with some extended boundary in  $C_{\mathcal{A}^\Delta}(V)$ , is smaller than  $\delta$ , provided that*

$$k \geq \frac{1}{\log_e(v+1)} \left[ v N r_{\max}^{W,V}(\Delta) + 6 \log_e \frac{evN}{6} + 6 \log_e \frac{l^2}{2l^\Delta \Delta} + \log_e c(\mathcal{A}, V, \Delta) + \log_e \frac{1}{\delta} \right], \quad (16)$$

where  $e$  is the natural logarithm base and  $v$  is any positive constant.

Note the increase in the size of the cover, due to the increased feature density in the extended boundary, which is reflected in the bound. A more careful analysis, which does

not rely directly on Lemmas 2 and 3, further improves (decreases) the bound (see the full version [10]).

### 3.4 Recognition Failures Due to the Combined Effect of Clutter and Similarities

Now, we combine the effects of clutter and object similarities, and integrate the partial expressions, derived above, into a uniform tool for evaluating the amount of data sufficient for recognition success. We consider images which contain both objects and noisy background, and are characterized by the following *mixed data model*.

**DEFINITION 2** (a mixed data model). *The features in the data set are determined randomly and independently according to the following piecewise uniform distribution:*

- In regions that belong to extended boundaries objects in the image, the density is  $f_b$ , while, in the rest of the image, the density is  $f_c < f_b = sf_c$ .
- The fraction of area occupied by extended boundaries is  $\alpha$ .
- The relative location of the objects in the scene is random, so that the objects “do not cooperate to cheat the recognition algorithm.” In particular, we assume that, if some hypothesized extended boundary  $V_t^\Delta$  maximally intersects with the extended boundary corresponding to a certain object in the image, denoted  $W_t^\Delta$ , then the nonoverlapping part of  $V_t^\Delta$ ,  $(V_t^\Delta \setminus W_t^\Delta = V_t^\Delta \cap \overline{W_t^\Delta})$ , intersects with other extended boundaries in at most  $\alpha' = \alpha - \frac{\text{Area}(W_t^\Delta)}{I^2}$  of its area. (The parameter  $\alpha'$  is the fraction of area occupied by extended boundaries of object instances different than  $W_t^\Delta$ .)

Invoking the constraint  $I^2 [\alpha \cdot f_b + (1 - \alpha) \cdot f_c] = 1$ , implies that

$$f_c = \frac{1}{I^2} \frac{1}{1 + \alpha(s-1)} \quad f_b = \frac{1}{I^2} \frac{s}{1 + \alpha(s-1)}.$$

Note that the scenarios considered above are special cases of the mixed model:  $s = 1$  corresponds to the uniform clutter model  $f_c = f_b - \frac{1}{I^2}$  and  $s \rightarrow \infty$  corresponds to the clutter free model  $f_c = 0$ ;  $f_b = \frac{1}{\alpha I^2}$ .

Let  $V, W_1, W_2, \dots$  be the objects that may appear in the scene. Let

$$r_{total} = \max \left[ r_{max}^V(\Delta, d_0), \max_i r_{max}^{W_i}(\Delta) \right]$$

be the *total similarity* parameter characterizing the worst case similarity. (As already mentioned, the similarity depends on the particular position of the object  $W_i$ . Therefore, the notation  $\max_i r_{max}^{W_i}(\Delta)$  may correspond either to a particular position attached to every object  $W_i$ , or to the maximal value over a range of allowed transformations. The results apply correspondingly either to a specific scene or to scenes which contains any instances of the objects  $\{W_i\}$  corresponding to this transformation range.)

The number of data features required for reliable localization in this most complicated scene is given by Theorem 1, stated in the beginning of this section. We now outline its proof.

**PROOF** (of Theorem 1—outline).

- 1) First, use the mixed model to bound the probability that a random data feature is included in some particular extended boundary. The bound, denoted  $p(s, \alpha, r_{total})$ , (see Theorem 1) depends the similarities of objects, but also on the image characterizations  $s$  and  $\alpha$ , and is called *effective similarity*.
- 2) Then, use Lemma 2 to find the size of the cover to the object instance class, which corresponds to the density specified by the mixed model.
- 3) Finally, use Lemma 3 to get the required number on the number of data features.  $\square$

Much like the similarity parameter in the case of clutterless scenes, the *effective similarity*,  $p(s, \alpha, r_{total})$ , quantifies the difficulty of the task for this more complicated and realistic scene model. For low saliencies, it is close to the probability  $2L^\Delta \Delta / I^2$ , corresponding to clutter-only scenes. When the saliency grows, the effective similarity increases and may achieve the highest similarity between the hypothesized object and some object in the scene. Interestingly, if the number of objects increases, but the number of feature points extracted from the scene stays constant, the effective similarity decreases (see Fig. 3 for an example). This happens because, in this case, the data features are divided between the objects and cannot work coherently to cheat the recognition algorithm, as happens in the case of clutterless scene, including only one object similar to the hypothesized one. The following corollary generalizes Theorem 1 and emphasizes the effect of the transformation class.

**COROLLARY 6** (to Theorem 1). *Consider the scene described by the mixed model and the localization task. Assume, however, that the hypothesized object instances are instances associated with either the Affine, Similarity, Euclidean, or Translations transformation classes. Then, the localization procedure is guaranteed, with confidence  $1 - \delta$ , not to pro-*

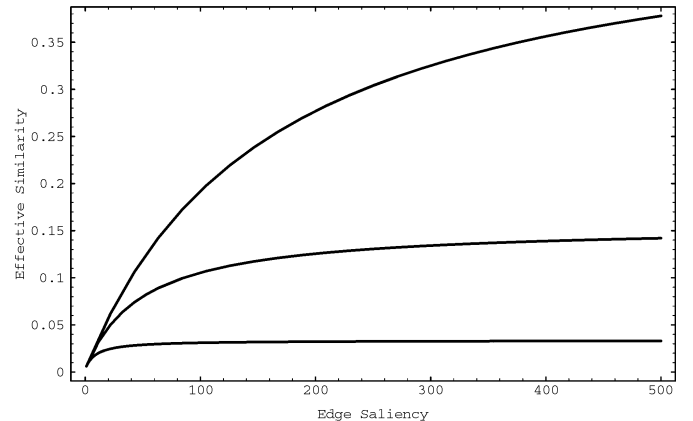


Fig. 3. The effective similarity that characterizes a localization task for which the total similarity is 0.5, plotted as a function of the boundaries' saliency and for several scenes occupied differently by objects: The upper curve represents a scene in which only one object exists (corresponding to  $\alpha$  value of 0.015) and the lower curves represent scenes which include other objects too ( $\alpha = 0.02$  (middle curve) and  $\alpha = 0.1$  (lowest curve)). Other parameters are  $L = 400$ ,  $I = 512$ ,  $\Delta = 2$ .

TABLE 1

Transformation class	$\kappa^*$	$\gamma^*$
Translations	2	$c(\mathcal{T}, V, \Delta) = 2 \cdot 6^3 (l + L^\Delta/2)^2 (L^\Delta)^2 l^{-4}$
Euclidean	3	$c(\mathcal{E}, V, \Delta) = \sqrt{2}\pi 6^3 (l + L^\Delta/2)^2 (L^\Delta)^3 l^{-6}$
Similarity	4	$c(\mathcal{S}, V, \Delta) = \pi 6^4 (l + L^\Delta/2)^2 (L^\Delta)^4 l^{-8}$
Affine	6	$c(\mathcal{A}, V, \Delta) = \frac{1}{2} \cdot 6^6 (l + L^\Delta/2)^2 (L^\Delta)^6 l^{-12}$

duce any false hypothesis, if

$$k \geq \frac{1}{\log_e(v+1)} \left[ vNp(s, \alpha, r_{total}) + \kappa^* \log_e \frac{evNs}{\kappa^*(1 + \alpha(s-1))} + \log_e \gamma^* + \log_e \frac{1}{\delta} \right], \quad (17)$$

where  $\kappa$  and  $\gamma$  are as specified in Table 1 and all other parameters are identical to those specified in Theorem 1.

#### 4 FROM LOCALIZATION TO RECOGNITION

A simple extension of our result would be to consider the general model-based recognition task, in which the hypotheses are instances of all objects in a given library  $\mathcal{L} = \{V_1, V_2, \dots, V_{|\mathcal{L}|}\}$ . Intuitively, more objects imply that the number of possible hypotheses also increases, which increases the probability to err. To keep the hypothesis reliable, the data subset size should therefore increase.

**CLAIM 7.** Let  $S$  be a set of  $N$  random features drawn in an  $l \times l$  image according to the mixed data model. Consider now a recognition procedure that relies on consistent data subsets of size  $k$  and tries to recognize an object from the library  $\mathcal{L} = \{V_1, V_2, \dots, V_{|\mathcal{L}|}\}$  under the affine transformation group  $\mathcal{A}$  with required localization precision  $d_0$ . Let  $r_{total}(V_i)$  be the total similarity parameter that characterizes the localization task of the  $i$ th object in  $\mathcal{L}$ . Then, the recognition procedure is guaranteed, with confidence  $1 - \delta$ , not to produce any false hypothesis, if

$$k \geq \frac{1}{\log_e(v+1)} \left[ vNp_{max}(s, \alpha, r_{total}) + 6 \log_e \frac{evNs}{6(1 + \alpha(s-1))} + \log_e \sum_i c(\mathcal{A}, V_i, \Delta) + \log_e \frac{1}{\delta} \right], \quad (18)$$

where  $c(\mathcal{A}, V_i, \Delta)$  is given in (10),  $v$  is any positive number, and

$$p_{max}(s, \alpha, r_{total}) = \frac{1}{l^2} \max_{t \in \mathcal{A}V_i \in \mathcal{L}} \left\{ \text{area}((V_i)_t^\Delta) \left[ 1 + \frac{r_{total}(V_i)(s-1)(1-\alpha)}{1 + \alpha(s-1)} \right] \right\}.$$

**PROOF (sketch).**  $p_{max}(s, \alpha, r_{total})$  is clearly an upper bound on the probability of a particular data feature to be inside the extended boundary of some particular instance of one of the objects in  $\mathcal{L}$ . A cover to the set of all objects' instances is a union of the covers associated with every individual object, and its size is simply the sum of their sizes. The rest is similar to the proof of Lemma 3.  $\square$

The claim predicts that, when the library size grows, the size of the required data subset grows only by a logarithmic factor. This is not surprising and was shown already in the context of clutterless scenes [9], [7]. Moreover, it is known by experiments that some recognition algorithms run in time proportional to the logarithm of the library size [16]. We have shown here that not only the computational effort, but also the amount of information required to succeed in the task, increases with logarithmic rate.

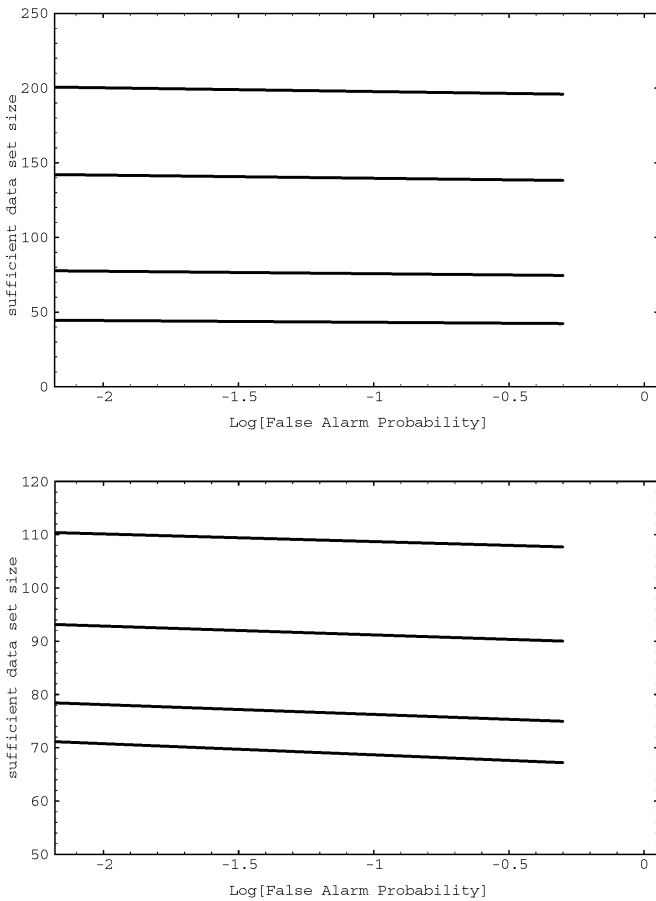


Fig. 4. The sufficient set size, as a function of the required reliability, for the affine transformation, and where the saliency is a parameter taking the values of 50, 20, 5, and 1 (top to bottom in the upper figure). The sufficient set size as a function of the required reliability, for saliency of 10, and for various transformations (affine, similarity, Euclidean, and translation)(top to bottom in the lower figure).

## 5 THE NECESSARY CONSISTENT SUBSET SIZE

In most of the paper, we derived an upper bound on the *sufficient* number of data features required as conclusive evidence for hypothesizing the presence of the object. In this section, we derive a lower bound, or an expression for a minimal data set size *necessary* for guaranteeing such high confidence hypothesis. In the next section, we derive another related figure, the maximal data set size which guarantees that the object in the scene is not missed.

To find the necessary subset size, consider a particular instance of one of the wrong objects (or a particular unacceptable instance of the true object) and observe that the number of data features inside the corresponding extended boundary is a binomial random variable. Then, the required necessary subset size is just the  $1 - \delta$  percentile, as lower subsets fall in this extended boundary more often than  $\delta$  and, therefore, are not reliable indicators for its presence in the scene. The tightest bound is obtained for a scene and object instance, maximizing the probability of a data feature to fall in the corresponding extended boundary. This maximal probability, which is just the effective similarity,  $p_0 = p(s, \alpha, r_{total})$ , and the total number of measurements,  $N$ , determine the distribution. Therefore, calculating the lower bound is simply specified by limiting the tail probability:

$$k_{necessary} = \sum_{i=k}^N \binom{N}{i} p_0^i (1-p_0)^{N-i} < \delta \quad (9)$$

## 6 MISS PROBABILITIES AND THE ROC CURVE

The probability of wrong hypothesis can be made arbitrarily small if the hypothesis is made only when the consistent number of data features is high enough and exceeds some threshold. Choosing a threshold which is too high, however, increases the likelihood of another type of recognition error: Missing an object instance while it is there. Therefore, an arbitrarily high threshold will also result in recognition failure.

The number of data features in the extended boundary of the correctly hypothesized object instance  $V_t$  is binomially distributed and is specified by the probability  $p_0$ , which is one in a clutterless, one object scene, but is lower in the general mixed model scene.

$$p_0 = \text{Area}[V_t^\Delta] \cdot f_b = \frac{\text{Area}[V_t^\Delta]}{f^2} \frac{s}{1 + \alpha(s-1)}$$

A perfect algorithm, which checks all hypotheses, does not miss the correct hypothesis with high probability  $1 - \delta$  if the consistent data sets size threshold by which an hypothesis is accepted is low enough. By the Chernoff bound, if the consistent set size,  $k$ , satisfies

$$k \leq Np_0 - \sqrt{3Np_0 \log \frac{1}{\delta}}, \quad (20)$$

then the probability of a miss error is not higher than  $\delta$  [14]. The miss probability is actually lower, because other acceptable hypotheses, which are close to  $V_t$ , are not missed either, with some nonzero probability. On the other hand,

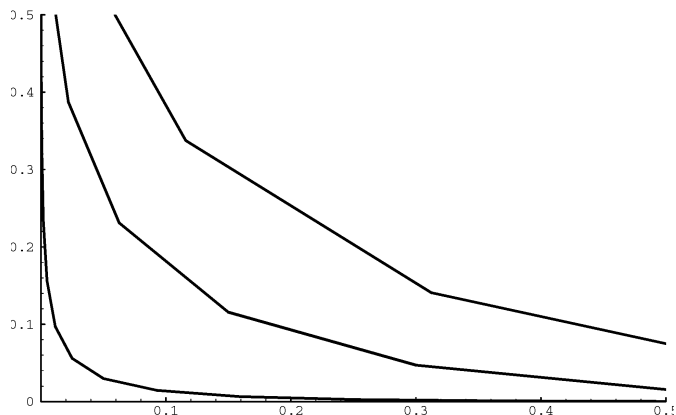


Fig. 5 The miss probability, plotted against a lower bound on the probability of false alarm for a recognition task associated with a total similarity parameter of 0.64. (The scene object is object (a) and the hypothesized instances are instances of object (c). The saliency was 10. See the simulation section.) The different curves are plotted for data feature set sizes of 100, 200, and 500, where increasing the amount of data corresponds to a better curve which is closer to the origin.

practical recognition systems do not necessarily consider the true instance of the hypothesized object as a candidate, but only some close instance out of some finite instance set. Therefore, their miss rate may be slightly higher. We expect, however, that, if the deviation is small, the instance space quantization is usually fine enough. (Note that, for such systems, the false alarm rate is lower.)

Plotting the false alarm probability against the miss probability gives the Receiver operating curve (ROC) commonly used for evaluating the performance achievable by decision systems (see an example related to recognition in [13]). Here, the consistent data sets size threshold by which a hypothesis is accepted serves as a parameter and is not apparent from the graph. Yet, the curve provides the information on the various trade-offs between the two error types that may be achieved by selecting the threshold. The curves illustrate another result of our analysis: the dependence of the recognition performance on the amount of data extracted from the image (see Fig. 6). Larger data sets put the ROC curve closer to the origin and increase performance.

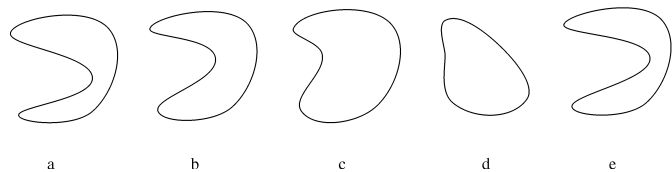


Fig. 6. The objects used for the simulations: The scene was based on either object a, b, c, or d, and the hypotheses were based on object e.

The curves of Fig. 6 actually show a lower bound on the false hypothesis probability (derived in the previous section) against the miss probability. Therefore, they should be considered as bounds: No recognition algorithm based on consistency, which operates on an image containing a specific number of data features, can do better than predicted by these curves.

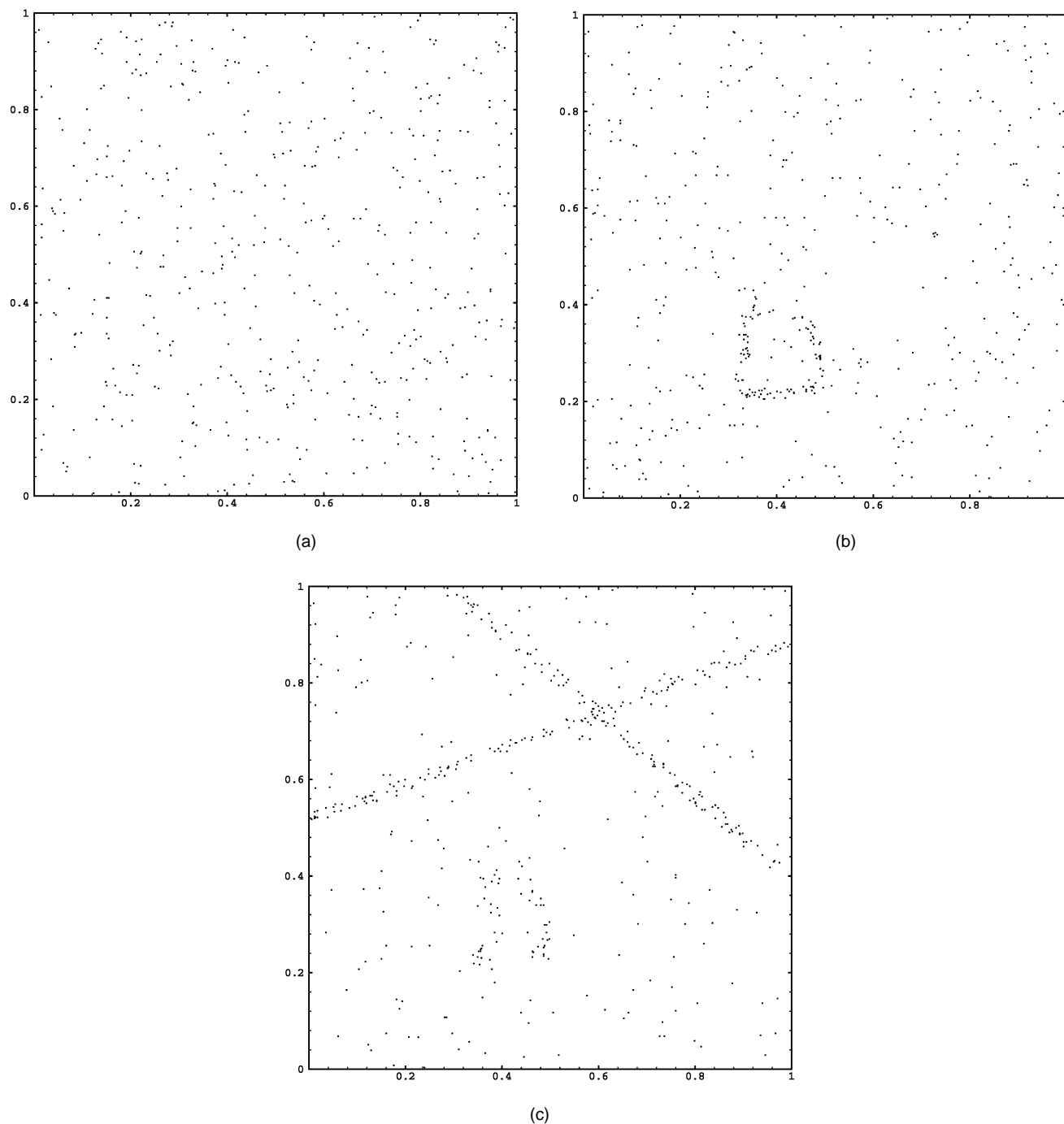


Fig. 7. Typical scenes that were used for the tests. Fig. 7a is associated with a saliency of one, meaning that the object is indistinguishable from the clutter. The two other scenes are associated with a saliency of 20, which makes the object (in Fig. 7b) or the object and the distractors (in Fig. 7c) stand out. (The image size is 512 by 512 and the measurements are associated with inaccuracy  $\Delta = 5$ .)

## 7 SIMULATION RESULTS

The results we developed in this paper should hold for every recognition scheme which accepts every hypothesis consistent with sufficiently large data features subset. Verifying this claim experimentally is difficult, because particular hypothesizing mechanisms (e.g., alignment) usually consider only a limited set of candidate hypotheses. As mentioned above, such procedures are characterized by a higher miss rate and lower false alarm. A notable exception is possible when the object is specified by a finite number of geometric primitives

(e.g., line segments). Then, the transformation space may be divided into a finite number of equivalence regions, thus leading to a finite time algorithm for evaluating the consistency of all possible hypotheses [17].

Here, we use one of the simplest recognition methods, a Hough transform, to test the bounds. This Hough algorithm is used for calculating the similarity and self-similarity parameters (see [7]). It is chosen because of its simplicity and availability, but mainly because, unlike some other simple algorithm, it scans all the transformation space (with some

TABLE 2

Scene Object Object	D	S	Effective Similarity	Confidence	predicted sufficient set size	predicted necessary set size	actual sufficient set size
a	no	1	0.012	0.99	29	13	21
a	no	10	0.094	0.99	104	64	65
a	no	50	0.307	0.99	265	177	190
b	no	10	0.079	0.99	92	55	56
b	no	50	0.253	0.99	225	150	158
c	no	10	0.071	0.99	87	50	51
c	no	50	0.227	0.99	210	137	142
d	no	10	0.049	0.99	69	37	41
d	no	50	0.146	0.99	151	93	100
c	no	10	0.071	0.999	90	55	56
c	no	10	0.071	0.99	87	50	51
c	no	10	0.071	0.9	84	44	45
c	no	50	0.227	0.99	210	137	142
c	1	50	0.152	0.99	145	96	89
c	2	50	0.114	0.99	119	75	68
d	no	1	0.012	0.9	30	21	23
d	no	10	0.067	0.9	83	42	48

quantization) without skipping any substantial region. As it is too slow and space demanding for testing the predictions related to the more complicated transformation classes, we have limited our simulations to the simpler translation and Euclidean classes. Yet, these results are enough to show, at least for these transformations, the validity of the derived upper bounds and also the gap that still exist between the predictions and the actual values.

The objects shown in Fig. 6 and the discrimination task were considered in the experiments. In all the tests, the scene was created by the objects a, b, c, or d, and the false hypothesis that some instance of the object e is in the scene was examined. For the Translation class, the similarities between the scene objects and the model are  $r_{max}^{a,e} = 0.888$ ,  $r_{max}^{b,e} = 0.71$ ,  $r_{max}^{c,e} = 0.64$ , and  $r_{max}^{d,e} = 0.40$ , respectively. We constructed synthetic scenes according to the mixed data model, each image containing one of these objects, and possibly additional straight distractors. Some typical images are shown in Fig. 7.

For every random scene, the data features obtained were used as an input to a Hough transform algorithm yielding a maximal vote for the translation associated with the in-

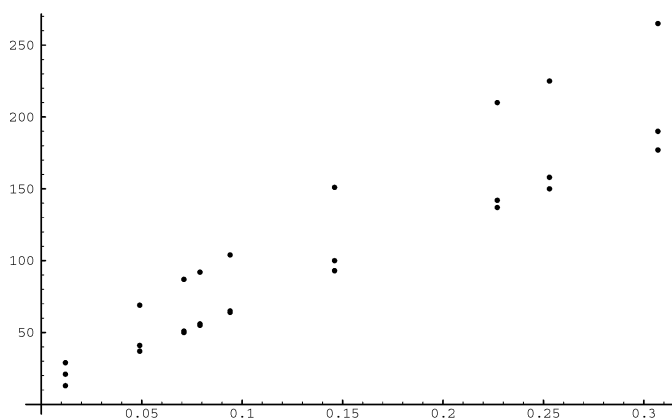


Fig. 8. The predicted bounds on the sufficient set size (upper and lower points) and the actual sufficient set size as determined by experiments (middle points), plotted against the effective similarity which characterizes the scene.

stance having the maximal consistency score with the data. We tested several cases corresponding to different scene objects saliencies and confidence values. For every case, the scene was created 5,000 times and the statistics of the maximal vote in the Hough transform were evaluated and compared with the predictions. A representative subset of the results is shown in Table 2, where the predicted (calculated) sufficient set size, the predicted (calculated) necessary set size, and the actual (measured) sufficient set size, as well as the effective similarity, are given as a function of the scene object, the number of distractors (D), and the saliency (S). (We used the slightly improved upper bound, described in the full version [10], to calculate the sufficient set size.)

The upper part of the table corresponds to cluttered images containing one object. Note that all bounds are valid for all the experimental situations tested. There is always a gap, however, between the upper bounds and the corresponding actual results, meaning that these bounds are not tight. On the other hand, the lower bounds are tighter and usually almost coincide with the actual measurements. Both the results and the predictions are also plotted in Fig. 8. It is apparent that it is not the similarity or the saliency alone which determine the sufficient subset size, but rather their cumulative effect as represented by the effective similarity parameter.

The second part of the table corresponds to the same object (c) and saliency (10) and examine the sufficient set size for different confidence values. (some lines are repeated in the table for easier reference.). Naturally, higher confidence requires larger data, but note that the difference is not so large relative to the difference related to similarity or saliency.

The third part of the table relates to scenes which include, besides the false object, other objects assumed to have a low similarity to the hypothesized object. In this case, these are simply straight lines (see Fig. 7). Note that their inclusion decreases both the predicted and the actual sets, as expected. Note also that, in this case, the lower bounds are pretty close but do not hold strictly. We attribute this small discrepancy to the random model specifying the effect of the distractors: According to this model, the

hypothesized object may overlap with high density areas that arise from both the object in the scene and from the distractors, implying that the effective similarity increases. If, however, the distractors are far from the scene object (as described in Fig. 7), then these contributions do not add and the actual effective similarity is lower, which results in lower discriminating data set size. Note that this discrepancy does not occur when the scene is cluttered uniformly.

The fourth and last part of the table describes some experiments with object instances corresponding to the Euclidean transformation class. This was, naturally, more difficult to simulate, due to the exponential growth of the simple Hough transform algorithm we used. Therefore, we conducted fewer tests (50 instead of 5,000). Still, the few results agree with the theoretical prediction developed above. We used the object *d* for creating the cluttered scene and tested all instances of object *e* against this data. The similarity between the scene object and the model was  $r_{max}^{d,e} = 0.564$ . Note that it is larger, as expected, than the same similarity for the translation case. This already implies that the set size, sufficient for discrimination, will be larger in the Euclidean case. Note, however, that our predictions, which take into account the effect of the transformation generality (via the cover size), imply that the discriminating set size will be even larger than the one attributed only to the increase in the similarity parameter. Our results indeed support this prediction, as is apparent from the table, where, for similar effective similarities, larger set sizes are required in the Euclidean case.

## 8 DISCUSSION

The research on object recognition provides many intuitive observations, such as "similar objects are harder to discriminate," "finer localization fails more often," "clutter makes localization more difficult," and "more objects in the candidate object library, require more data but not much more." These observations are supported by many experiments. Here, we analyze the information required for recognition tasks, and provide analytic and quantitative support for the above intuitive observations.

The analysis considers different kinds of recognition failures and quantify the combined effect of

- 1) the amount of clutter in the image,
- 2) the transformation class,
- 3) the presence of similar objects in the scene, and the self-symmetry of the hypothesized object itself,
- 4) the accuracy of the data,
- 5) the localization precision required, and
- 6) the confidence we would like to have in our hypotheses.

We provide expressions for the amount of data extracted from the image, that, if consistent with some object hypotheses, provides a sufficiently reliable evidence for the existence of this object in the imaged scene. The results apply to every algorithm that uses such a consistent data set size as a criterion for accepting an hypothesis.

Note that the claims assume that the objects expected in the scene are known. More precisely, we just have to know the similarity between *V* and the scene object which is different from it but is the most similar to it. This is natural to

demand, because, otherwise, very similar objects to *V* may be present, leading inevitably to a lot of false hypotheses. If this information is not known in advance, then our claims should be interpreted as conditional ones, that is, as relations between the total similarity that characterizes the task and the sufficient data set size.

The results we provide are useful, for example, for setting the threshold when designing a new algorithm, and for analyzing reported results by comparing them to the theoretical bounds. To do that, one should first characterize the scene using the  $\alpha$  and  $s$  parameters, which are influenced by the lighting and clutter conditions. Then, the objects involved should be considered and the similarity characterizing the task should be calculated from their similarities and self-similarities. Finally, Theorem 1 should be used to determine the threshold which guarantees reliable hypotheses. Besides this straightforward application, the bounds give insight to the recognition process and explicitly show how the amount of information required for success is affected by the various scene and task parameters. This allows us to infer the performance of a recognition procedure that was tested in a particular setting, when conditions are changed.

Our bounds are not always tight. This is attributed to the worst case assumptions we take in the derivation and to its noncompromise rigorness. In [7], [10], we used more sophisticated analysis to get tighter bounds. We expect, however, that further improvements should be possible.

An even more interesting subject is the sufficient size of data subset for data features that carry additional information, such as the curve slope, the curvature, etc. A preliminary work, which considers the effect of slope in a different framework, is presented in [18], [9].

## REFERENCES

- [1] W.E.L. Grimson and D.P. Huttenlocher, eds., special double issue on the Interpretation of 3D Scenes, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 10, Oct. 1991, and vol. 14, no. 2, Feb. 1992.
- [2] *Geometric Invariance in Computer Vision*, J.L. Mundy and A.P. Zisserman, eds. MIT Press, 1992.
- [3] W.E.L. Grimson and D.P. Huttenlocher, "On the Verification of Hypothesized Matches in Model-Based Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 12, pp. 1,201-1,213, Dec. 1991.
- [4] W.E.L. Grimson, D.P. Huttenlocher, and D.W. Jacobs, "A Study of Affine Matching with Bounded Sensor Error," *Proc. European Conf. Computer Vision*, pp. 291-306, 1992.
- [5] K.B. Sarachik and W.E.L. Grimson, "Gaussian Error Models for Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 400-406, 1993.
- [6] S.J. Maybank, "Probabilistic Analysis of the Application of the Cross Ratio to Model Based Vision," *Int'l J. Computer Vision*, vol. 16, pp. 5-33, 1993.
- [7] M. Lindenbaum, "Bounds on Shape Recognition Performance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 666-680, July 1993.
- [8] S. Ben-David and M. Lindenbaum, "Localization vs. Identification of Semi-Algebraic Sets," *Proc. Sixth ACM Conf. Computational Learning Theory*, pp. 327-336, 1993.
- [9] M. Lindenbaum and S. Ben-David, "Applying vc-Dimension Analysis to Object Recognition," *Proc. European Conf. Computer Vision*, pp. 239-240, 1994, to appear in *J. Machine Intelligence and Vision*.
- [10] M. Lindenbaum, "On the Amount of Information Required for Object Recognition," CIS Report 9329, Technion, Nov. 1993

(revised July 1995). A shorter version appeared in *Proc. Int'l Conf. Pattern Recognition*, pp. 726-729, 1994.

- [11] T.M. Breuel, "Higher-Order Statistics in Object Recognition," *Computer Vision and Pattern Recognition*, pp. 707-708, 1993.
- [12] A. Amir and M. Lindenbaum, "Grouping Based Non-Additive Verification," CIS Report 9518, Computer Science Dept., Technion, 1996, to appear *IEEE Trans. Pattern Analysis and Machine Intelligence*.
- [13] K.B. Sarachik, "The Effect of Gaussian Error in Object Recognition," *Proc. Image Understanding Workshop*, pp. 1,269-1,280, 1994.
- [14] T. Hagerup and C. Rub, "A Guided Tour of Chernoff Bounds," *Information Processing Letters*, vol. 33, pp. 305-308, 1989.
- [15] W.J. Rucklidge, "Locating Objects Using the Hausdorff Distance," *Proc. Int'l Conf. Computer Vision*, pp. 457-464, 1995.
- [16] P.G. Gottschalk, J.L. Turney, and T.N. Mudge, "Efficient Recognition of Partially Visible Objects Using a Logarithmic Complexity Matching Technique," *Int'l J. Robotic Research*, vol. 8, no. 6, pp. 110-131, 1989.
- [17] T.A. Cass, "Polynomial Time Object Recognition in the Presence of Clutter Occlusion, and Uncertainty," *Proc. European Conf. Computer Vision*, pp. 834-842, 1992.
- [18] A. Rudshstein and M. Lindenbaum, "Qualifying the Performance of Feature-Based Recognition," *Proc. Int'l Conf. Pattern Recognition*, pp. 35-39, 1996.



**Michael Lindenbaum** received his BSc, MSc, and DSc degrees from the Department of Electrical Engineering at the Technion, Israel, in 1978, 1987, and 1990, respectively. He did his postdoctoral research at the NTT Basic Research Labs in Tokyo. Since October 1991, he has been with the Department of Computer Science at the Technion. His main research interest is computer vision, especially statistical analysis of object recognition and grouping processes.