Predicting Object Recognition Performance Under Data Uncertainty, Occlusion and Clutter*

Michael Boshra and Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, California 92521
{michael,bhanu}@vislab.ucr.edu

Abstract

We present a novel method for predicting the performance of an object recognition approach in the presence of data uncertainty, occlusion and clutter. The recognition approach uses a vote-based decision criterion, which selects the object/pose hypothesis that has the maximum number of consistent features (votes) with the scene data. The prediction method determines a fundamental, optimistic, limit on achievable performance by any vote-based recognition system. It captures the structural similarity between model objects, which is a fundamental factor in determining the recognition performance. Given a bound on data uncertainty, we determine the structural similarity between every pair of model objects. This is done by computing the number of consistent features between the two objects as a function of the relative transformation between them. Similarity information is then used, along with statistical models for data distortion, to estimate the probability of correct recognition (PCR) as a function of occlusion and clutter rates. The method is validated by comparing predicted PCR plots with ones that are obtained experimentally.

1 Introduction

Model-based object recognition is a central problem in image analysis. It can be defined as follows: Given a set of model objects and scene data provided by a sensor observing one of these objects, the objective is to determine identity and pose of the scene object. Object recognition involves extracting features from the scene data and finding consistent correspondence between scene features and those of a model object. Accordingly, performance of this process depends on properties of both the scene data (e.g., measurement error, missing and spurious features) and the model objects (e.g., articulation of object parts, objects having similar subparts). Modeling all these factors is a challenge for predicting the recognition performance.

We present a method for predicting the probability of correct recognition (PCR) by considering: 1) *Scene-Data Factors:* data uncertainty (due to measurement error), occlusion (missing scene-object features), and clutter (spurious scene features), and *2) Model Factors:* the structural similarity between model objects. The model similarity factor is fundamental for determining the recognition performance. Intuitively, the probability of failing to recognize an object, in a distorted scene, is directly proportional to the degree of similarity between this object and the rest of the model objects. We assume that model objects and scene data are represented by 2-D point features, where each feature is represented by its positional information. Further, we assume that the decision criterion is vote-based; i.e., the object/pose hypothesis with the maximum number of consistent features is selected as the valid one.

Related research efforts consider model similarity, data uncertainty, and occlusion [1], or model similarity, data uncertainty, and clutter [2]. Other relevant efforts address the problem of discriminating an object from random clutter (e.g., [3]). The problem of predicting PCR as a function of data uncertainty, occlusion, clutter and model similarity, which is the focus of this paper, has not been adequately addressed in the field.

2 Approach

Our performance prediction problem can be stated as follows: Given 1) a set of model objects, $\mathcal{M} = \{\mathcal{M}_i\}$, 2) statistical models for scene-data distortion (uncertainty, occlusion and clutter), and 3) a class of applicable transformations, $\mathcal{T}$ (e.g., translation, rigid, affine), our objective is to predict the PCR plot as a function of occlusion and clutter rates, assuming a fixed amount of data uncertainty. The proposed

*This work was supported in part by DARPA/AFOSR grant F49620-97-1-0184; the contents and information do not reflect positions or policies of the U.S. Government.

0-8186-8821-1/98 $10.00 © 1998 IEEE

method can be outlined as follows. *Firstly*, for each model object, $\mathcal{M}_i \in \mathcal{M}$, we compute the structural similarity between $\mathcal{M}_i$ and every model object $\mathcal{M}_j \in \mathcal{M}$ (j may be equal to i). The similarity between $\mathcal{M}_i$ and $\mathcal{M}_j$, which is simply the number of consistent features, is computed as a function of the transformation of $\mathcal{M}_j$ relative to $\mathcal{M}_i$. *Secondly*, given model similarity information, obtained in the first step, and specific occlusion and clutter rates, we compute the probability of correctly recognizing each model object. Taking the average PCR for all model objects and repeating the process for a variety of occlusion and clutter rates, we can predict the PCR plot.

The major issues involved in our method are discussed in the remainder of this section.

1) Data Distortion Models: We model data distortion as follows: *1) Data Uncertainty:* The actual location of a scene feature is assumed to be uniformly distributed within a circle of radius ϵ, which is centered at the estimated feature location. *2) Occlusion:* we assume that each feature subset of $\mathcal{M}_i$ is equally likely to be occluded as any other feature subset that is of the same size. *3) Clutter:* Clutter features are assumed to be uniformly distributed within some area surrounding the object.

2) Model Similarity: The similarity between $\mathcal{M}_i$ and $\mathcal{M}_j$ is a function of the relative transformation between them. Let us refer to an *instance* of $\mathcal{M}_j$ at location $\tau \in \mathcal{T}$ relative to $\mathcal{M}_i$ as $\mathcal{M}_j^\tau$. The similarity between $\mathcal{M}_i$ and $\mathcal{M}_j^\tau$, $S_\epsilon(\mathcal{M}_i, \mathcal{M}_j^\tau)$, is defined as *the expected number of votes that $\mathcal{M}_j^\tau$ would get, given uncertain scene data of $\mathcal{M}_i$*. We determine $S_\epsilon(\mathcal{M}_i, \mathcal{M}_j^\tau)$ as follows. Feature/feature similarity, $FFS_\epsilon(F_{ik}, F_{jl}^\tau)$, between features $F_{ik} \in \mathcal{M}_i$ and $F_{jl}^\tau \in \mathcal{M}_j^\tau$ is defined as *the probability that an uncertain measurement of F_{ik} is consistent with F_{jl}^τ*. In our work, a pair of scene and model features are considered *consistent*, if they lie within a distance of ϵ. Thus,

$$FFS_\epsilon(F_{ik}, F_{jl}^\tau) = \frac{A(R_\epsilon(F_{ik}) \cap R_\epsilon(F_{jl}^\tau))}{\pi\epsilon^2}$$

where $R_\epsilon(F)$ is the *uncertainty region* associated with feature F, a circle of radius ϵ that is centered at F, and $A(R)$ is the area of region R. Thus, the similarity between F_{ik} and F_{jl}^τ is proportional to the intersection area of the respective uncertainty regions, as illustrated in Figure 1. The similarity between $\mathcal{M}_i$ and $\mathcal{M}_j^\tau$ can be defined as follows:

$$S_\epsilon(\mathcal{M}_i, \mathcal{M}_j^\tau) \approx \lfloor (\sum_k \sum_l FFS_\epsilon(F_{ik}, F_{jl}^\tau)) + 0.5 \rfloor.$$

The above equation is approximate because it assumes one-to-one correspondence between *similar* scene and model features, which are those with overlapping uncertainty regions (see Figure 2).

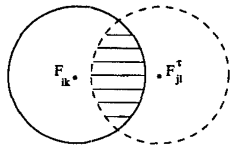


Figure 1: The similarity between F_{ik} and F_{jl}^τ is proportional to the intersection area of the respective uncertainty regions (the shaded region).

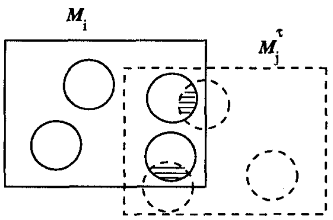


Figure 2: An illustration of one-to-one correspondence between similar features in $\mathcal{M}_i$ and $\mathcal{M}_j^\tau$.

3) Effects of Data Distortion and Model Similarity on Recognition Performance: Figure 3 shows a schematic diagram of $\mathcal{M}_i$ and $\mathcal{M}_j^\tau$, showing similar and dissimilar features in both objects. The intersection region denotes the similar features in $\mathcal{M}_i$ ($\mathcal{M}_{ij}$) or $\mathcal{M}_j^\tau$ ($\mathcal{M}_{ji}^\tau$), while the other regions denote dissimilar features in $\mathcal{M}_i$ ($\mathcal{M}_{i/j} = \mathcal{M}_i/\mathcal{M}_{ij}$), and $\mathcal{M}_j^\tau$ ($\mathcal{M}_{j/i}^\tau = \mathcal{M}_j^\tau/\mathcal{M}_{ji}^\tau$). To simplify both the presentation in this section and the probabilistic analysis in the next section, we assume that the intersection region corresponds to $S_\epsilon(\mathcal{M}_i, \mathcal{M}_j^\tau)$ features that are *fully* similar, i.e., $FFS_\epsilon(\cdot,\cdot) = 1$.

A distorted image of $\mathcal{M}_i$ is falsely interpreted as $\mathcal{M}_j^\tau$, if $\mathcal{M}_j^\tau$ gets more votes than $\mathcal{M}_i$. The probability of false recognition increases as: 1) Features in $\mathcal{M}_{i/j}$, which distinguish $\mathcal{M}_i$ from $\mathcal{M}_j^\tau$, start to get occluded, 2) clutter features happen to coincide with those in $\mathcal{M}_{j/i}^\tau$, 3) the size of the intersection area increases, due

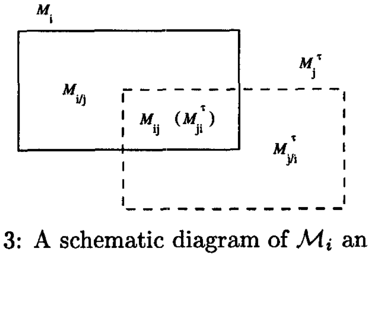


Figure 3: A schematic diagram of $\mathcal{M}_i$ and $\mathcal{M}_j^\tau$.

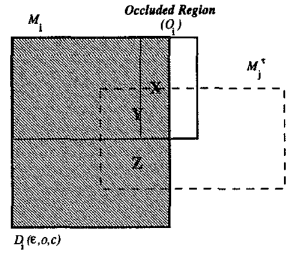


Figure 4: Vote variables for $\mathcal{M}_i$ and $\mathcal{M}_j^\tau$, shown at the center of their corresponding regions.

to the similarity between the two objects, as well as the increase in the amount of data uncertainty. Notice that, in our vote-based scheme, neither occlusion nor clutter in the intersection area has any effect on the performance of recognition, since the vote difference between $\mathcal{M}_i$ and $\mathcal{M}_j^\tau$ would stay the same.

3 Performance Estimation

In this section, we estimate the probability of correctly recognizing $\mathcal{M}_i$ in a distorted image, $\mathcal{D}_i(\epsilon, o, c)$, where ϵ is the radius of data uncertainty, o is the number of occluded features of $\mathcal{M}_i$, and c is the number of clutter features. We start by assuming that $\mathcal{M}_i$ can be falsely recognized as only one object instance, $\mathcal{M}_j^\tau$, and then we consider the general case.

1) Single-Instance Case: Let $\mathcal{F}_\epsilon(\mathcal{M}_k; \mathcal{D}_l)$ be the set of features of image $\mathcal{D}_l$ that are consistent with hypothesis $\mathcal{M}_k$, and $V_\epsilon(\mathcal{M}_k; \mathcal{D}_l)$ the number of votes for $\mathcal{M}_k$, given $\mathcal{D}_l$. If there is one-to-one correspondence between consistent scene and model features, then $V_\epsilon(\mathcal{M}_k; \mathcal{D}_l) = \mid \mathcal{F}_\epsilon(\mathcal{M}_k; \mathcal{D}_l) \mid$. Further, let $\mathcal{O}_i$ be the set of occluded features of $\mathcal{M}_i$ in $\mathcal{D}_i(\epsilon, o, c)$ (notice that $\mid \mathcal{O}_i \mid = o$). The probability of misinterpreting $\mathcal{D}_i(\epsilon, o, c)$ as $\mathcal{M}_j^\tau$ can be written as

$$\begin{aligned}&\Pr[\mathcal{M}_j^\tau; \mathcal{D}_i(\epsilon, o, c)] = \\ &\qquad \Pr[V_\epsilon(\mathcal{M}_j^\tau; \mathcal{D}_i(\epsilon, o, c)) \geq V_\epsilon(\mathcal{M}_i; \mathcal{D}_i(\epsilon, o, c))]. (1)\end{aligned}$$

It can be shown that the votes for $\mathcal{M}_i$ and $\mathcal{M}_j^\tau$ are

$$\begin{aligned}V_\epsilon(\mathcal{M}_i; \mathcal{D}_i(\epsilon, o, c)) &= \mid \mathcal{M}_i \mid - o + X, \text{ and} \\ V_\epsilon(\mathcal{M}_j^\tau; \mathcal{D}_i(\epsilon, o, c)) &= Y + Z\end{aligned}$$

where X, Y, and Z are random variables that are defined as follows: 1) $X = V_\epsilon(\mathcal{O}_i; \mathcal{D}_i(\epsilon, o, c))$, the number of votes for $\mathcal{O}_i$ due to coincidence with clutter features of $\mathcal{D}_i(\epsilon, o, c)$, 2) $Y = V_\epsilon(\mathcal{M}_j^\tau; \mathcal{F}_\epsilon(\mathcal{M}_i, \mathcal{D}_i(\epsilon, o, c)))$, the number of votes for $\mathcal{M}_j^\tau$ due to the similarity with $\mathcal{M}_i$, and 3) $Z = V_\epsilon(\mathcal{M}_j^\tau; \mathcal{D}_i(\epsilon, o, c)) - Y$, the number of votes for $\mathcal{M}_j^\tau$ which are not due to the similarity with $\mathcal{M}_i$. These variables are illustrated in Figure 4.

We determine the probability distribution functions (PDF's) of X, Y, Z, and the votes for $\mathcal{M}_i$ and $\mathcal{M}_j^\tau$ based on the statistical data-distortion models outlined in the previous section. Let us first determine the PDF describing X and Z. Given m model features and n clutter features, the probability that u model features are consistent with clutter features is bounded by $[G_U(u; m, n, 0), G_U(u; m, n, 1)]$ where

$$\begin{aligned}&G_U(u; m, n, p) = \\ &\qquad C(m, u) P(n, u) \left(\frac{\pi\epsilon^2}{I}\right)^u \left(\frac{I - (m - pu)\pi\epsilon^2}{I}\right)^{n-u},\end{aligned}$$

$C(a, b) = \frac{a!}{(a-b)!\, b!}$, $P(a, b) = \frac{a!}{(a-b)!}$, and I is the clutter area. Since we are interested in determining an optimistic estimate of PCR, we use the upper and lower bounds for describing X and Z, respectively.

Next, let us determine the PDF of Y. Since we are assuming that all feature subsets of $\mathcal{M}_i$, of the same size, are equally likely to be occluded, it can be shown that Y is described by a hypergeometric distribution,

$$H_Y(y; S_\epsilon(\mathcal{M}_i, \mathcal{M}_j^\tau), u, \mid \mathcal{M}_i \mid - u)$$

where $u = V_\epsilon(\mathcal{M}_i; \mathcal{D}_i(\epsilon, o, c))$ and

$$H_U(u; n, a, b) \quad = \quad \frac{C(a, u) C(b, n - u)}{C(a + b, n)}.$$

It is easy to show that the PDF of the vote count for $\mathcal{M}_i$ is

$$\Pr[V_\epsilon(\mathcal{M}_i; \mathcal{D}_i(\epsilon, o, c)) = u] = G_X(u - \mid \mathcal{M}_i \mid + o; o, c, 1), \quad (2)$$

while that for $\mathcal{M}_j^\tau$ is

$$\begin{aligned}&\Pr[V_\epsilon(\mathcal{M}_j^\tau; \mathcal{D}_i(\epsilon, o, c), u) = v] = \\ &\qquad \sum_y H_Y(y; S_\epsilon(\mathcal{M}_i, \mathcal{M}_j^\tau), u, \mid \mathcal{M}_i \mid - u) \times \\ &\qquad G_Z(v - y; \mid \mathcal{M}_j^\tau \mid - y, c - u - \mid \mathcal{M}_i \mid + o, 0) \quad (3)\end{aligned}$$

where $V_\epsilon(\mathcal{M}_j^\tau; \mathcal{D}_i(\epsilon, o, c), u)$ is the number of votes for $\mathcal{M}_j^\tau$, assuming that $\mathcal{M}_i$ has u votes.

From (1), (2) and (3), we determine the probability of misinterpreting $\mathcal{D}_i(\epsilon, o, c)$ as $\mathcal{M}_j^\tau$:

$$\begin{aligned}\Pr[\mathcal{M}_j^\tau; \mathcal{D}_i(\epsilon, o, c)] &= \sum_u \Pr[V_\epsilon(\mathcal{M}_i; \mathcal{D}_i(\epsilon, o, c)) = u] \times \\ &\qquad \sum_{v \geq u} \Pr[V_\epsilon(\mathcal{M}_j^\tau; \mathcal{D}_i(\epsilon, o, c), u) = v].\end{aligned}$$

2) General Case: Now, we consider the general case, where a distorted scene of $\mathcal{M}_i$ can be misinterpreted as any $\mathcal{M}_j^\tau$, $\forall \mathcal{M}_j \in \mathcal{M}$, and $\forall \tau \in \mathcal{T}$, except for a small window around the origin of $\mathcal{T}$ when $j = i$. Clearly this case is much more difficult, since $\mathcal{M}_i$ can be misinterpreted as an infinite number of possible object instances. Discretization appears to be a necessary approach in order to facilitate the PCR estimation process. One possibility is to discretize the transformation space $\mathcal{T}$ at some fine resolution, and consider object instances at the sampling points only. The main problem with this approach is the vote dependence between adjacent instances, which would significantly complicate the analysis. In order to overcome this problem, we select object instances that correspond to the *peaks* of the similarity function. The rationale behind this choice can be stated as follows: 1) Peaks are generally not very close to each other and so the vote independence assumption (or more accurately, the conditional vote independence assumption, see below) becomes reasonable. 2) The probability that an off-peak instance, $\mathcal{M}_j^{\tau+\delta}$, gets more votes than the neighboring peak instance, $\mathcal{M}_j^{\tau}$, is small. This is because of the high overlapping between the uncertainty regions of $\mathcal{M}_j^{\tau}$ and $\mathcal{M}_j^{\tau+\delta}$, for smaller δ's, and the lesser number of similarity votes (votes due to similarity with $\mathcal{M}_i$) for $\mathcal{M}_j^{\tau+\delta}$, for larger δ's. The consequence of using only peak instances in the analysis is producing an optimistic PCR.

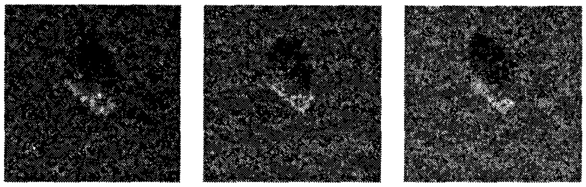
Let $\{\mathcal{M}_j^{\tau}\}$ be the set of peak object instances for model object $\mathcal{M}_i$. The probability of correctly recognizing $\mathcal{M}_i$ can be written as follows:

$$\Pr[\mathcal{M}_i;\mathcal{D}_i(\epsilon,o,c)] = \sum_u \Pr[V_\epsilon(\mathcal{M}_i;\mathcal{D}_i(\epsilon,o,c)) = u] \times \prod_j (1 - \Pr[V_\epsilon(\mathcal{M}_j^{\tau};\mathcal{D}_i(\epsilon,o,c),u) \geq u)]. \quad (4)$$

Notice that we assume *conditional* vote independence in the above equation. Since the probability of misinterpretation to a particular instance depends on instance size, and number of similar features only, we can rewrite (4) as

$$\Pr[\mathcal{M}_i;\mathcal{D}_i(\epsilon,o,c)] = \sum_u \Pr[V_\epsilon(\mathcal{M}_i;\mathcal{D}_i(\epsilon,o,c)) = u] \times \prod_n \prod_s (1 - P(u;n,s))^{N_i(n,s)} \quad (5)$$

where $P(u;n,s) = \Pr[V_\epsilon(\mathcal{M}_j^{\tau};\mathcal{D}_i(\epsilon,o,c),u) \geq u]$, assuming that $|\ \mathcal{M}_j^{\tau}\ | = n$, $S_\epsilon(\mathcal{M}_i,\mathcal{M}_j^{\tau}) = s$, and $N_i(n,s)$ is the number of peak object instances of size n and similarity size s. The PCR can be estimated by evaluating (5) for each model object and taking the average. A linear approximation of (5) is used in the implementation, to avoid computing it for each object separately.



(a) T72 (b) BMP2 (c) BTR

Figure 5: Examples of target views.

4 Experimental Results

In this section, we consider the task of recognizing targets in Synthetic Aperture Radar (SAR) images, in order to experimentally validate our prediction method.

1) Model Data: Each model target is represented by a number of SAR views which sample its signature at a specific depression angle (θ_d), and a variety of azimuth angles. The model database consists of views corresponding to three targets: T72 (231 views), BMP2 (233 views) and BTR (233 views) at $\theta_d = 17°$. Examples of these views are shown in Figure 5. Each view is treated as an independent object for recognition purposes. In our case, the space of applicable transformations is 2-D translation in the image plane [4]. Scattering centers, peaks in the image, are used as point features for recognition. These peaks are extracted by comparing the value of each pixel with its eight neighbors. We have chosen the strongest 30 scattering centers to represent both model and scene data. Since we are considering a fixed number of scattering centers, the occlusion and clutter rates in an image are always the same.

2) Test Data: Test data are obtained by selecting a number of target views, and introducing distortion to these views as follows: *1) Data Uncertainty:* The location of each peak is randomly perturbed such that the distance between the new peak and the original one is smaller than a desirable radius of data uncertainty, ϵ. The new peak is then quantized to coincide with a pixel location. *2) Occlusion:* A feature is randomly eliminated along with a number of its nearest neighbors, depending on the desired occlusion rate. *3) Clutter:* A number of features, depending on the desired clutter rate, are randomly generated within the target bounding box.

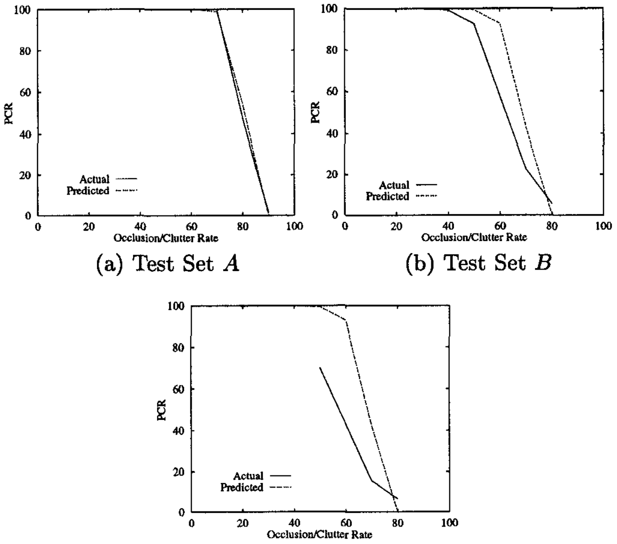
We have constructed three test sets, A, B and C, which are described in Table 1. Note that due to the depression angle variation between the model data (17°) and the test set C (15°), data distortion is naturally introduced in the original test views of C. The occlusion/clutter rate and ϵ for these original views are estimated to be about 50% and 1 pixel, respectively.

Table 1: Description of test data sets A, B, and C.

Set	Test views	ϵ	Occlusion/Clutter
A	697 model views	0	20%, 30%, ..., 90%
B	697 model views	1.0	20%, 30%, ..., 80%
C	413 views, $\theta_d = 15°$	1.0	50%, 60%, ..., 80%

3) Recognition System: Our recognition system is based on geometric hashing. The relative distances, along image axes (range, cross range), between every pair of model scattering centers are used to map the pair into a tuple in the hash table, which consists of target type, azimuth and location of the reference scattering center. At run-time, pairs of scattering centers, extracted from the scene data, are used to generate votes for the identity, azimuth and location of the scene target. A pair of scene and model scattering centers are considered consistent, if they are in the same location, in case $\epsilon = 0$, or if they are four-neighbors, in case $\epsilon = 1.0$. The generated votes are accumulated in a 4-D vote table, and the hypothesis corresponding to the entry with the maximum number of votes is selected. For test sets A and B, the generated hypothesis is considered correct, if the target type and azimuth are the same as those of the scene target, and the scene/model relative location is not more than ϵ along each of the image axes. For test set C, due to the nature of the problem and the data, we have allowed the azimuth to be within ±4° and the relative location to be within ±5 pixels along each image axis. The algorithm examines almost all of the target, azimuth and discretized relative location space, and so we consider its performance to be close to optimal.

4) Results: Figure 6 shows the PCR plots for the three test sets. From these plots, we observe: 1) In case $\epsilon = 0$, Figure 6(a), the predicted plot virtually coincides with the actual one. This is because, in such a case, the similarity function is composed of weighted impulse functions. Accordingly, peak instances, which are considered for PCR estimation, represent all possible candidates for misinterpretation, thus leading to a very accurate estimate. 2) In case $\epsilon = 1.0$, Figures 6(b) and 6(c), our method provides an optimistic PCR estimate. This is expected, since, as explained in the previous section, peak instances no longer represent all possible candidates for misinterpretation (which are infinite in this case). Another, less obvious, reason for the optimism of PCR estimation is the difference between the consistency region used in the experiments, four-neighbors, and the one assumed in the analysis, a circle of radius ϵ. The fact that the area of the former is about 60% larger than that of the later increases the clutter votes for object instances, thus degrading performance and further deviating actual PCR plot from predicted one.



(a) Test Set A

(b) Test Set B

(c) Test Set C

Figure 6: Actual and predicted probability of correct recognition plots for the three test data sets.

5 Conclusions

We have presented a novel method for predicting the performance of object recognition as a function of data uncertainty, occlusion, clutter and model similarity. Validity of the method has been demonstrated by comparing predicted PCR plots with those that are obtained experimentally using SAR data.

References

[1] M. Lindenbaum. Bounds on shape recognition performance. *IEEE Trans. PAMI*, 17(7), 1995.

[2] M. Lindenbaum. An integrated model for evaluating the amount of data required for reliable recognition. *IEEE Trans. PAMI*, 19(11), 1997.

[3] W. E. L. Grimson and D. P. Huttenlocher. On the verification of hypothesized matches in model-based recognition. *IEEE Trans. PAMI*, 13(12), 1991.

[4] B. Bhanu and G. Jones III. Performance characterization of a model-based SAR target recognition system using invariants. In *SPIE Conf. Algorithms for Synthetic Aperture Radar Imagery IV*, volume 3070, 1997.