REPEATED SAMPLING TO IMPROVE CLASSIFIER ACCURACY*

Jiangying Zhou Daniel Lopresti

Matsushita Information Technology Laboratory
Panasonic Technologies, Inc.
Two Research Way
Princeton, NJ 08540 USA
[jz,dpl]@mitl.research.panasonic.com

ABSTRACT

In statistical pattern recognition, the Bayes Risk serves as a reference – a limit of excellence that cannot be surpassed. In this paper, we show that by relaxing the assumption that the input be sampled only once, a classification system can be built that beats the Bayes error bound. We present a detailed analysis of the effects of repeated sampling, including proofs that it always yields a net improvement in recognition accuracy for common distributions of interest. Upper and lower bounds on the net improvement are also discussed. We conclude by giving preliminary experimental results that illustrate the applicability of this approach.

INTRODUCTION

A fundamental problem in pattern recognition is to take an unidentified object and associate it with one of a set of pre-defined classes according to the measurement of some number of its physical attributes. It is well known that the error rate for any statistical classifier based on a specific collection of attributes, or *feature set*, is lower-bounded by the Bayes Risk [1, 3]. In this paper, we show that by relaxing a basic assumption – that the input be sampled only once – a classification system can be built that beats the Bayes error bound. This result is not just a theoretical curiosity, but appears to have practical applications in real-world recognition problems.

According to the Bayes theorem, the design of a statistical classifier is dictated by the characteristics of the *a priori* class probabilities and by the conditional probability distributions of the measured features for each class. Once the distributions of these random variables are known, the optimal classification boundaries are determined by the Bayes decision rule. Errors arise when the distributions for different classes overlap (*e.g.*, Figure 1). In the traditional case, such mistakes are unavoidable; the classifier is "optimal" in the sense that it minimizes this base error rate.

In a previous paper, we introduced a methodology that reduces the residual error rate in optical character recognition (OCR) by sampling the input repeatedly and combining the results through a novel voting scheme [6]. We observed that between 20% and 40% of the OCR errors were eliminated when we simply scanned a page three times and applied *consensus sequence voting* on the output from a particular OCR package. We speculate that when the performance of a recognition process is very high (*e.g.*, 99% or higher), a significant portion of the remaining errors arise from "unlucky" random fluctuations in the input data. In this paper, we present a formal analysis of this effect, showing that better performance – beyond the limit of the Bayes error bound – can be achieved by exploiting the small variations inherent in observed measurements. We also present preliminary experimental results that illustrate the application of this approach to a specific problem in machine vision.

PRELIMINARIES

A pattern recognition system can be viewed as consisting of three parts: a set of pattern classes, an observation space, and a decision mechanism. Pattern classes represent abstract categories from which objects are drawn. Examples include symbols over a given alphabet (*e.g.*, $\{a, b, c, \ldots\}$), editing gestures made with a pen, components to be assembled by a robot, etc. We denote the set

*Presented at the *IAPR Workshop on Machine Vision Applications*, Kawasaki, Japan, December 1994.

of pattern classes as $\mathbf{C} = \{C_1, C_2, \ldots, C_m\}$. The observation space, also abstract, is a vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ representing information that can be extracted from objects (*e.g.*, color, texture, length of a substructure, angle of a curve). For a specific instance of an object, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $x_i \in X_i$, represents the set of values $\mathbf{X}$ takes on. During the recognition process, these quantities are measured, and a class assignment is made by the decision mechanism.

In reality, it is not possible to determine a single, "true" feature vector $\mathbf{x}$ for an object. Rather, $\mathbf{x}$ is sampled via a stochastic process. In some sense, the innate value is *hidden* from direct observation. For example, we know that football fields are supposed to be 100 yards long. Hence, if $X_1 = \textit{length}$ is a feature of interest, then we would expect $x_1 = \textit{100 yards}$ for a particular football field. However, in the case of real football fields, we are likely to see a series of slightly different values, even when measuring the same field twice. Hence, any assessment of $\mathbf{X}$ is inevitably embedded in some randomness, and the recognition system can only obtain an approximation of the value $\mathbf{x}$.

To make this distinction clear, we employ "hat" notation $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ to designate the observations returned by the stochastic process. In the example above, we might have $\hat{x}_1 = \textit{100.1 yards}$, *99.7 yards*, *100.2 yards*, ..., as a succession of measurements. The set of all possible $\hat{\mathbf{x}}$'s is denoted by $\hat{\mathbf{X}}$. We further let M represent the stochastic process. For a given object γ with innate feature vector $\mathbf{x}$, each output from M is a random vector defined on a probability space by the conditional probability:

$$Pr(\mathbf{X} = \hat{\mathbf{x}} \,|\, \gamma) = \mathbf{Pr}(\mathbf{X} \textit{ is measured as } \hat{\mathbf{x}} \,|\, \gamma), \qquad (1)$$

or the corresponding probability density function $p_M(\hat{\mathbf{x}} \,|\, \gamma)$.

In this paper, we model the observation process for a given class as an additive perturbance $\hat{\mathbf{X}} = \mathbf{X} + N$, where N represents random "noise." We denote the probability density function governing the perturbance as $p_N(\hat{x} \,|\, C_i)$, and let $p(\mathbf{x} \,|\, C_i)$ signify the probability density function of the hidden random vector $\mathbf{X}$, taken over all objects in class C_i. The distribution of $\hat{\mathbf{X}}$ is described by the conditional probability density function $p(\hat{\mathbf{x}} \,|\, C_i)$. In the following discussion, the term *hidden distribution* refers to $p(\mathbf{x} \,|\, C_i)$, the term *primary distribution* refers to $p(\hat{\mathbf{x}} \,|\, C_i)$, and the term *secondary distribution* refers to $p_M(\hat{\mathbf{x}} \,|\, \gamma)$. A *sample* is the outcome of an observation process, while *repeated sampling* refers to observations made of the same input.

When only a single observation $\hat{\mathbf{x}}$ for a given object is used, the decision rules are determined by the conditional probability functions $p(\hat{\mathbf{x}} \,|\, C_i)$ for all classes $C_i \in C$.

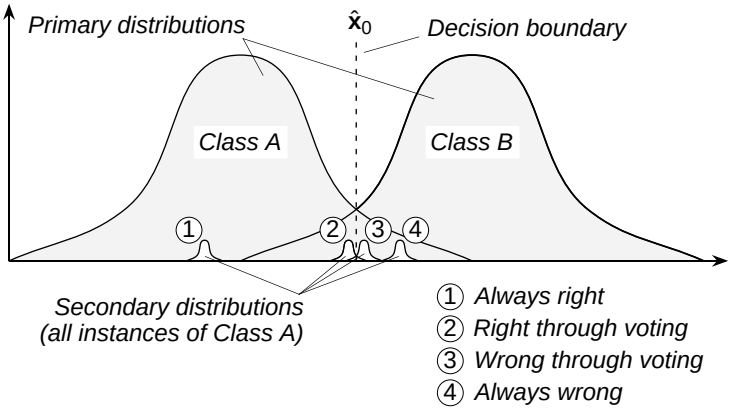

Figure 1: Primary and secondary probability distributions for a two-class recognition problem.

The rules that minimize the overall expected error rate are known as the Bayes decision rules. We use $Pr(\epsilon \,|\, C_i)$ to represent the mis-classification rate for class C_i. The Bayes risk is a weighted average of the mis-classification rates for all classes: $P_e = \sum_{i=1}^{i=M} Pr(\epsilon \,|\, C_i) P(C_i)$. Note that the Bayes risk is the smallest error rate possible for a given feature set under the indicated observation process. In traditional classifier theory, any improvement beyond this requires changing the feature set and/or the observation process. Even if it were possible for the observation process to be perfect (*i.e.*, noiseless), mistakes might still be inevitable: two objects from different classes possessing the same innate feature vector $\mathbf{x}$ are effectively indistinguishable. We call such errors the *intrinsic* errors of a recognition system and its associated feature set.

AN APPROACH FOR BEATING THE BAYES BOUND

In Bayes classifier design, the discussion is centered on "single use" decision situations. That is, a single observation is made of an input and then a decision is reached based on that observation. The premise of our technique for improving on the Bayes bound lies in the fact that the underlying physical attributes of an object can be sampled more than once (*e.g.*, a page of text can be scanned several times). Since the measurement of a field datum is a random variable, the outcome of each sample is potentially different.

For purposes of illustration, suppose we have two classes, A and B, with conditional probability distributions as shown in Figure 1. The Bayes decision boundary is $\hat{\mathbf{x}} = (x_0)$. Now assume that we are presented with a particular instance α from class A, and that the measured fea-

ture has a secondary Gaussian-type distribution $p_M(\hat{\mathbf{x}} \mid \alpha)$ with mean μ_α. It is clear that if μ_α lies close to the decision boundary, a given observation $\hat{\mathbf{x}}$ may fall on the wrong side, resulting in a classification error. The probability this event happens is $\int_{x_0}^{\infty} p_M(\hat{\mathbf{x}} \mid \alpha)\, d\hat{\mathbf{x}}$.

However, if several measurements are taken, the majority of them should fall on the proper side of the decision boundary (see case (2) in Figure 1). In other words, recognition is made more reliable in spite of individual failures by taking the consensus of repeated samples of the input. Conversely, if the mean falls on the wrong side of the boundary, $\mu_\alpha > x_0$, the voting scheme may actually do more harm than good (case (3) in Figure 1). Intuitively, though, it should be evident from the shape of the primary distribution for class A that this approach is likely to work more often than not. In the following sections, we show that there is always a net improvement in recognition accuracy for common distributions of interest.

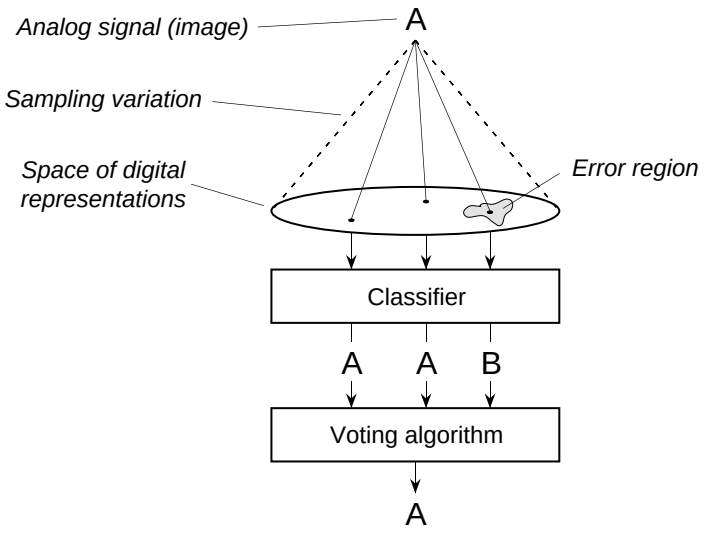

Figure 2: Pattern recognition using repeated sampling.

Figure 2 presents an overview of the approach. Note that the classifier in the figure is still limited by the Bayes bound. However, the performance of the system as a whole can be better than this, as we have noted. To demonstrate this more concretely, we performed a simulation of the two-class problem illustrated in Figure 1. We chose means for the secondary distributions based on initial Gaussian distributions with $\mu_A = -2.0$, $\mu_B = 2.0$, and $\sigma = 1.0$. We then generated observation samples using secondary Gaussian distributions with σ's of 0.10, 0.20, and 0.30. As shown in Figure 3, the voting system's advantage over the "optimal" classifier ranged from 1.13% to 16.02%. The Bayes Risk in the simulation varied between 0.023 and 0.028, corresponding to initial recognition accuracies of 97.2% to 97.7%. These results seem consistent with the experimental OCR data, cited earlier, that originally motivated our investigation.

Samples (Voters)	Improvement Over Bayes Risk		
	$\sigma = 0.10$	$\sigma = 0.20$	$\sigma = 0.30$
3	1.13%	5.24%	10.12%
5	1.52%	6.55%	13.09%
7	1.53%	7.18%	14.40%
9	2.10%	7.65%	15.44%
11	2.19%	7.77%	16.02%

Figure 3: Simulation results showing the effects of repeated sampling.

ANALYSIS I – GAUSSIAN NOISE DISTRIBUTIONS

In this section we present a theoretical analysis of the improvement in the error rate achieved by repeated sampling for a particular class of noise distributions: the perturbance is independent of the hidden distribution $p(\mathbf{x} \mid C_i)$ and has a probability density function that is a zero-mean Gaussian.

Again, a two-class, one-dimensional problem is considered. Let $p(\mathbf{x} \mid A)$ and $\mathcal{N}(0, \sigma_A)$ denote the probability functions for the hidden and noise distributions for class A, and $p(\mathbf{x} \mid B)$ and $\mathcal{N}(0, \sigma_B)$ denote the hidden and noise distributions for class B, respectively. With no loss of generality, we assume that the Bayes decision boundary between the classes is located at $\hat{\mathbf{x}} = 0$. Thus, the classifier assigns label A to an object if $\hat{\mathbf{x}}$ is less than 0; otherwise the label B is assigned.

For a particular object $\alpha \in A$, say that the hidden value of $\mathbf{X}$ is $\mathbf{x}_\alpha$. Then the probability that a single sample falls in the region $\hat{\mathbf{x}} < 0$, denoted P_α, is

$$P_\alpha = Pr(\mathbf{x}_\alpha + N < 0) = \frac{1}{\sqrt{2\pi}\,\sigma_A} \int_{-\infty}^{0} e^{-\frac{(\hat{\mathbf{x}} - \mathbf{x}_\alpha)^2}{2\sigma_A^2}}\, d\hat{\mathbf{x}}$$

and the probability that $\hat{\mathbf{x}} > 0$ is, of course, $1 - P_\alpha$. We say that object α is correctly recognized by the Bayes classifier with probability P_α.

It is obvious that if $\mathbf{x}_\alpha < 0$, then $P_\alpha > 0.5$. Now suppose that $2m + 1$ independent observations of α are made. The probability that a majority of them fall on the side $\hat{\mathbf{x}} < 0$

is characterized by the equation:

$$\Phi(P_\alpha, m) = \sum_{i=m+1}^{2m+1} \left(\begin{array}{c} 2m+1 \\ i \end{array} \right) P_\alpha^i (1 - P_\alpha)^{2m+1-i}$$

for $m = 1, 2, \ldots$. It is easily demonstrated that $\Phi(P_\alpha, m) > P_\alpha$ whenever $0.5 < P_\alpha < 1$. Consequently, if majority voting is used, there is an increased chance α will be classified properly. Moreover, the probability that α is correctly recognized can be shown to approach 1 as $m \to \infty$. This result implies that if we have $\mathbf{x}_\alpha < 0$ for a given α, voting can lead to perfect recognition, whereas the Bayes classifier makes an error with probability $1 - P_\alpha$. Over all α's, the asymptotic improvement voting can bring about for class A is

$$C(A) = \int_{-\infty}^{0} p(\mathbf{x} \mid A) \left[\frac{1}{\sqrt{2\pi}\,\sigma_A} \int_0^\infty e^{-\frac{(\hat{\mathbf{x}}-\mathbf{x})^2}{2\sigma_A^2}}\, d\hat{\mathbf{x}} \right] d\mathbf{x} \quad (2)$$

On the other hand, if the hidden value lies on the wrong side of the decision boundary, $\mathbf{x}_\alpha > 0$, voting is likely to produce the incorrect answer, while the Bayes classifier might recognize α correctly (*i.e.*, if the observation $\hat{\mathbf{x}}$ happens to land on the right side, $\hat{\mathbf{x}} < 0$). The probability this kind of "cross-over" occurs is

$$\tilde{P}_\alpha = \frac{1}{\sqrt{2\pi}\,\sigma_A} \int_{-\infty}^{0} e^{-\frac{(\hat{\mathbf{x}}-\mathbf{x}_\alpha)^2}{2\sigma_A^2}}\, d\hat{\mathbf{x}}$$

Since $\tilde{P}_\alpha < 0.5$, we have $\Phi(\tilde{P}_\alpha, m) < \tilde{P}_\alpha$, hence voting decreases the chances of α being properly classified.

We refer to the situation where the Bayes classifier is right but voting returns the wrong result as *voting damage*. The overall damage induced by the voting scheme is upper-bounded by

$$D(A) = \int_{0}^{\infty} p(\mathbf{x} \mid A) \left[\frac{1}{\sqrt{2\pi}\,\sigma_A} \int_{-\infty}^0 e^{-\frac{(\hat{\mathbf{x}}-\mathbf{x})^2}{2\sigma_A^2}}\, d\hat{\mathbf{x}} \right] d\mathbf{x} \quad (3)$$

The net asymptotic improvement in recognition accuracy, $\Delta(A)$, is then defined as the difference between Equations (2) and (3):

$$\Delta(A) \equiv C(A) - D(A)$$

This can be expressed as

$$\Delta(A) = \frac{1}{2} \int_0^\infty [p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A)] \left[1 - \mathrm{erf}\left(\frac{\mathbf{x}}{\sigma_A \sqrt{2}} \right) \right] d\mathbf{x}$$

where

$$\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

See [7] for further details.

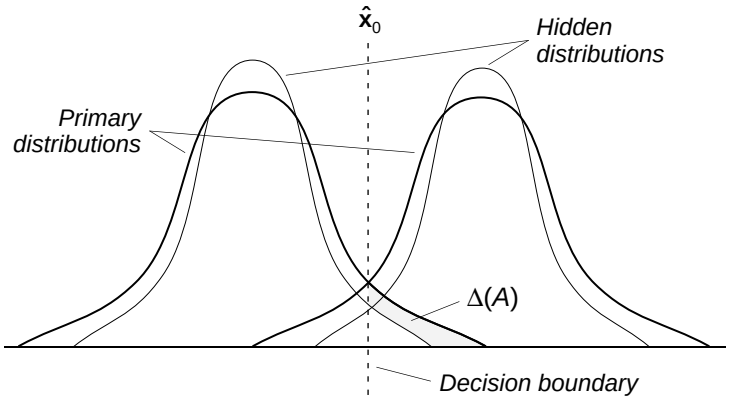

Figure 4: Net improvement in recognition accuracy, $\Delta(A)$.

PROPERTIES OF $\Delta(A)$ FOR GAUSSIAN NOISE

In this section, we present two theorems regarding the net improvement in recognition accuracy, $\Delta(A)$. The first expresses $\Delta(A)$ as a function of the Bayes and intrinsic error rates. The second shows that $\Delta(A) > 0$ (*i.e.*, repeated sampling always reduces the number of errors) for common hidden distributions of interest.

Theorem 1 *Let N be a Gaussian noise process. The net improvement $\Delta(A)$ due to repeated sampling equals the Bayes mis-classification rate minus the intrinsic error rate of the recognition process:*

$$\Delta(A) = Pr(\epsilon \mid A) - \int_0^\infty p(\mathbf{x} \mid A)\, d\mathbf{x} \quad (4)$$

Proof: Owing to space limitations, we refer the reader to [7] for the proof of this theorem.

This theorem implies that when the hidden distributions for classes A and B are separable (*i.e.*, if $\int_0^\infty p(\mathbf{x} \mid A) = 0$), $\Delta(A)$ equals the Bayes Risk. In other words, we achieve error-free recognition for class A.

Theorem 2 *Let N be a Gaussian noise process. For common hidden distributions of interest, repeated sampling always yields a net improvement in recognition accuracy. That is, $\Delta(A) > 0$.*

Proof: The proof is divided into three cases. First we show that the theorem is true for all "bell-shaped" (*e.g.* , Gaussian) distributions. Then we prove that the result holds for arbitrary distribution functions under certain reasonable conditions.

Case 1 – Bell-Shaped Distributions.

A bell-shaped function $f(x)$ is a non-negative function satisfying the following property: $f(x_1) > f(x_2)$ if $|x_1 - \mu| < |x_2 - \mu|$, where μ is the highest (extreme) point of the function. Now, suppose distribution function $p(\mathbf{x} \mid A)$ is bell-shaped with $\mu < 0$. Then we have $|-\mathbf{x} - \mu| < |\mathbf{x} - \mu|$ for all $\mathbf{x} > 0$, hence $p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A) > 0$. From the definition of the error function $\mathrm{erf}(x)$, it should be clear that $1 - \mathrm{erf}\left(\mathbf{x}/(\sigma_A\sqrt{2})\right)$ is positive for all $\mathbf{x} > 0$. Therefore, $\Delta(A) > 0$.

For the other two cases, we need to assume the following:

$$\int_{-c}^{0} p(\mathbf{x} \mid A)\, d\mathbf{x} - \int_{0}^{c} p(\mathbf{x} \mid A)\, d\mathbf{x} > 0 \qquad \forall c > 0 \qquad (5)$$

The term $\int_{-c}^{0} p(\mathbf{x} \mid A)\, d\mathbf{x}$ corresponds to the density of the intrinsic value $\mathbf{X}$ in the interval $[-c, 0]$, the neighborhood of size c to the left of the decision boundary (*i.e.*, *inside* the region for class A). Similarly, the term $\int_{0}^{c} p(\mathbf{x} \mid A)\, d\mathbf{x}$ is the density of $\mathbf{X}$ in the interval $[0, c]$, the same-sized neighborhood to the right of the decision boundary (*i.e.*, *outside* the region for class A). The condition states that the density of $\mathbf{X}$ is always higher on the correct side of the boundary than on the incorrect side. Intuitively, we can see why this should be necessary: otherwise "cross-overs" (*e.g.*, case (3) in Figure 1) will happen more frequently, meaning voting may do more harm than good.

Case 2 – Finite Support.

Suppose that the hidden distribution $p(\mathbf{x} \mid A)$ has finite support. That is, $p(\mathbf{x} \mid A) = 0$ for all $|x| > \Omega$, where Ω is a fixed positive value. This is a reasonable assumption since real-world systems are finite. In this case, $\Delta(A)$ can be written as

$$\Delta(A) = \frac{1}{2}\int_{0}^{\Omega} [p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A)] \left[1 - \mathrm{erf}\left(\frac{\mathbf{x}}{\sigma_A\sqrt{2}}\right)\right] d\mathbf{x}$$

By the Mean-Value Theorem, there exists a value $0 < \omega < \Omega$ such that

$$\begin{aligned} \Delta(A) &= \frac{1}{2}(1 - \mathrm{erf}(0)) \int_{0}^{\omega} [p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A)]\, d\mathbf{x} \\ &= \frac{1}{2}\int_{0}^{\omega} [p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A)]\, d\mathbf{x} \end{aligned}$$

Since our assumption implies that $\int_{0}^{c} p(-\mathbf{x} \mid A)\, d\mathbf{x} - \int_{0}^{c} p(\mathbf{x} \mid A)\, d\mathbf{x} > 0$ for all $c > 0$ (Equation 5), we have $\Delta(A) > 0$.

Case 3 – Infinite Support.

Lastly, we consider the case where $p(\mathbf{x} \mid A)$ has infinite support. We first show that repeated sampling performs no worse than the original classifier (*i.e.*, $\Delta(A) \geq 0$).

Consider the sequence of real-valued functions $f_1, f_2, \ldots$ where

$$f_m \equiv \frac{1}{2}\int_{0}^{m} [p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A)] \left[1 - \mathrm{erf}\left(\frac{\mathbf{x}}{\sigma_A\sqrt{2}}\right)\right] d\mathbf{x}$$

Obviously, we have $\lim_{m\to\infty} f_m = \Delta(A)$. Again, from the Mean-Value Theorem, we know that for each f_m there exists a value $0 < \omega_m < m$ such that

$$f_m = \frac{1}{2}\int_{0}^{\omega_m} [p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A)]\, d\mathbf{x}$$

Thus, using Equation 5, we have $f_m > 0$. Therefore, $\Delta(A) \geq 0$.

Finally, we can show there is always a net improvement (*i.e.*, $\Delta(A) > 0$) if we make one more assumption: for any given σ_A, there exists at least one real value $r > 0$ such that

$$\mathcal{G}_r \equiv \frac{1}{2}\int_{r}^{\infty} [p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A)] \left[1 - \mathrm{erf}\left(\frac{\mathbf{x}}{\sigma_A\sqrt{2}}\right)\right] d\mathbf{x} \geq 0$$

When this is the case, $\Delta(A)$ can be written as

$$\Delta(A) = \mathcal{F}_r + \mathcal{G}_r$$

where

$$\mathcal{F}_r = \frac{1}{2}\int_{0}^{r} [p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A)] \left[1 - \mathrm{erf}\left(\frac{\mathbf{x}}{\sigma_A\sqrt{2}}\right)\right] d\mathbf{x}$$

From Equation 5 we know that $\mathcal{F}_r > 0$. Hence $\Delta(A) > 0$. □

It can be demonstrated that the final condition is satisfied by most hidden distributions of interest [7].

ANALYSIS II – ARBITRARY SYMMETRIC NOISE DISTRIBUTIONS

The two theorems in the previous section can be extended to more general classes of observation perturbances. In fact, Theorems 1 and 2 remain true for any noise process with a symmetric probability density function $p_N(\hat{\mathbf{x}} \mid A)$ that is independent of the hidden distribution.

Theorem 3 *Let N be a random noise process with a symmetric distribution function $p_N(\hat{\mathbf{x}} \mid A)$ independent of the distribution of $\mathbf{X}$. The net improvement $\Delta(A)$ due to repeated sampling equals the Bayes mis-classification rate minus the intrinsic error rate of the recognition process:*

$$\Delta(A) = Pr(\epsilon \mid A) - \int_{\mathbf{x}_0}^{\infty} p(\mathbf{x} \mid A)\, d\mathbf{x}$$

Proof: The proof is similar to the proof of Theorem 2. The details are given in [7].

Theorem 4 *Let N be a random noise process with a symmetric distribution function $p_N(\hat{\mathbf{x}} \mid A)$ independent of the distribution of $\mathbf{X}$. If the following two conditions are satisfied:*

1. *For all $c > 0$,*

$$\int_{\mathbf{x}_0-c}^{\mathbf{x}_0} p(\mathbf{x} \mid A)\, d\mathbf{x} - \int_{\mathbf{x}_0}^{\mathbf{x}_0+c} p(\mathbf{x} \mid A)\, d\mathbf{x} > 0$$

2. *For any given $p_N(\hat{\mathbf{x}} \mid A)$, there exists at least one real value $r > 0$ such that*

$$\int_{\mathbf{x}_0+r}^{\infty} (p(-\mathbf{x} \mid A) - p(\mathbf{x} \mid A)) \left[\int_{\mathbf{x}}^{\infty} p_N(\hat{\mathbf{x}} \mid A)\, d\hat{\mathbf{x}}\right] d\mathbf{x} \geq 0$$

then repeated sampling always yields a net improvement in recognition accuracy. That is, $\Delta(A) > 0$.

Proof: The proof of this theorem can be found in [7].

BOUNDS ON $\Delta(A)$

In this section, we provide estimates for upper and lower bounds on the net improvement in recognition accuracy, $\Delta(A)$. We start by giving a very general upper bound in the case that $p_N(\hat{\mathbf{x}} \mid A)$ is fixed and symmetric.

Theorem 5 *For any fixed symmetric density function $p_N(\hat{\mathbf{x}} \mid A)$, $\Delta(A) \leq \frac{1}{2}$.*

Proof: The proof of this theorem can be found in [7].

One scenario where the maximum $\Delta(A)$ is achieved arises when the distribution of the hidden random variable $\mathbf{X}$ is an impulse function centered very close to the decision boundary, *i.e.*, at $\mathbf{x} = \hat{\mathbf{x}}_0 - \epsilon$. It can be proved that as $\epsilon \to 0$, we have $\Delta(A) \to 0.5$.

When both $p_N(\hat{\mathbf{x}} \mid A)$ and $p(\mathbf{x} \mid A)$ are Gaussian distributions, a lower bound can be shown. Suppose $\mathcal{N}(0, \sigma_A)$ and $\mathcal{N}(-\mu, \sigma_S)$ are the density functions for $p_N(\hat{\mathbf{x}} \mid A)$ and $p(\mathbf{x} \mid A)$, respectively. Since $\hat{\mathbf{X}} = \mathbf{X} + \mathbf{N}$ is the sum of two independent Gaussian random variables, it follows that $\hat{\mathbf{X}}$ is also a Gaussian random variable with mean $E\{\mathbf{X}+\mathbf{N}\} = -\mu$ and standard deviation $D\{\mathbf{X}+\mathbf{N}\} = \sigma_A^2 + \sigma_S^2$. In other words,

$$p(\hat{\mathbf{x}} \mid A) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_A^2 + \sigma_S^2}}\ e^{-(\hat{\mathbf{x}}+\mu)^2/2(\sigma_A^2+\sigma_S^2)}$$

Assuming the Bayes decision boundary is $\hat{\mathbf{x}} = 0$, by Theorem 1 we can write $\Delta(A)$ as

$$\begin{aligned}\Delta(A) &= Pr(\epsilon \mid A) - \int_0^{\infty} p(\mathbf{x} \mid A)\, d\mathbf{x} \\ &= \frac{1}{2}\left[\mathrm{erfc}\left\{\frac{\mu}{\sqrt{2(\sigma_A^2 + \sigma_S^2)}}\right\} - \mathrm{erfc}\left\{\frac{\mu}{\sigma_S\sqrt{2}}\right\}\right]\end{aligned}$$

where $\mathrm{erfc}() \equiv 1 - \mathrm{erf}()$.

When μ and σ_S are both fixed and the perturbance σ_A is small, *i.e.*, $\sigma_A \ll \sigma_S$, the expression can be expanded as Taylor series at $\frac{\mu}{\sigma_S\sqrt{2}}$ and we obtain:

$$\Delta(A) \geq \frac{\mu}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2\sigma_S^2}} \left(\frac{\sqrt{\sigma_A^2 + \sigma_S^2} - \sigma_S}{\sigma_S\sqrt{\sigma_A^2 + \sigma_S^2}}\right)$$

Figure 5 shows the lower bound as a function of σ_A computed at $\sigma_S = 1.0$, $\mu = 2.0$ (solid curve). The actual net improvement due to repeated sampling when there are 50 voters is shown as the marked curve in the same figure. For small σ_A, the bound is quite tight.

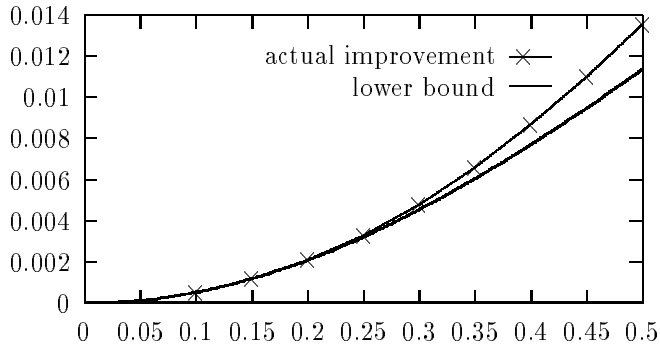

Figure 5: Net improvement in recognition accuracy, actual vs. lower bound.

EXPERIMENTAL RESULTS

In this section, we present some preliminary experimental data that demonstrate the applicability of repeated sampling.

As we noted previously, our OCR results show that a significant percentage (*i.e.*, 20-40%) of the residual errors in an otherwise accurate system can be corrected through repeated sampling. The extent to which this approach can help with other recognition problems is a subject we are currently investigating. Here we describe some early results that illustrate the degree of random fluctuation in the input data to a simple machine vision application.

The problem we have chosen to examine is deciding whether a given coin is showing *heads* or *tails*. This is a relatively challenging recognition task due to the highly-reflective, sculpted surface of the coin (a U.S. 1 cent piece, in our case). Figure 6 gives one of the test images from our experiments. Our goals are two-fold: (1) to verify the inherent randomness of the input process, and (2) to test the usefulness of repeated sampling in a real, albeit contrived, application.

Since our focus is on repeated sampling, we have elected to use a fairly "generic" recognition procedure. First, a segmentation algorithm based on morphological operations is employed to break the image into subregions containing individual coins. Once the coins are identified, we compute from each image a set of moments as described in [5]. This set is invariant with respect to size, position (translation), rotation, and reflection. The moments are taken as our features and provided as input to a simple linear classification algorithm. In a preliminary experiment, we used 120 coin images (60 heads and 60 tails) as a training set, and then tested using a different set of 240 coin images (120 heads and 120 tails). The overall recognition rate was 87%.

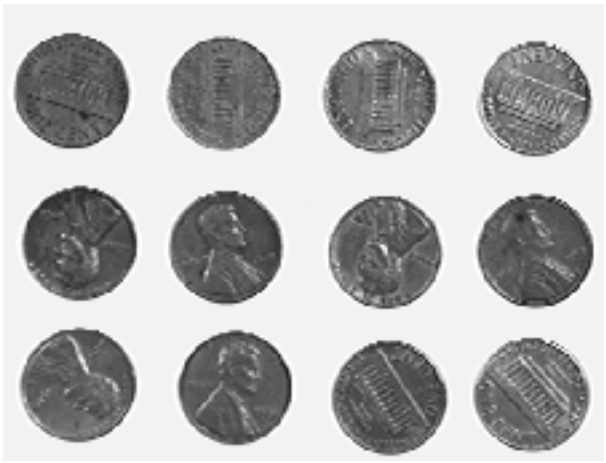

Figure 6: Sample image of 12 coins.

To see the effects of the perturbance induced by the imaging process, we sampled each coin three times. These three snapshots were taken in rapid succession using a Panasonic GP-MF200 camera without changing any settings or moving the coins. For each coin, we computed the mean of the feature vectors extracted from the three snapshots as well as the maximum variance between the mean and the vectors. Figure 7 shows the results calculated from the first two moments using 30 coins randomly chosen from the test set. Each circle represents the maximum feature variance for a particular coin. The straight line running through the plot is the decision boundary used by our classifier. As the figure illustrates, the three snapshots yield significant variation in the computed feature vectors; of the 25 coins represented in the region depicted, five cross the boundary. The potential impact on recognition results, especially for feature vectors near the decision boundary, is quite clear.

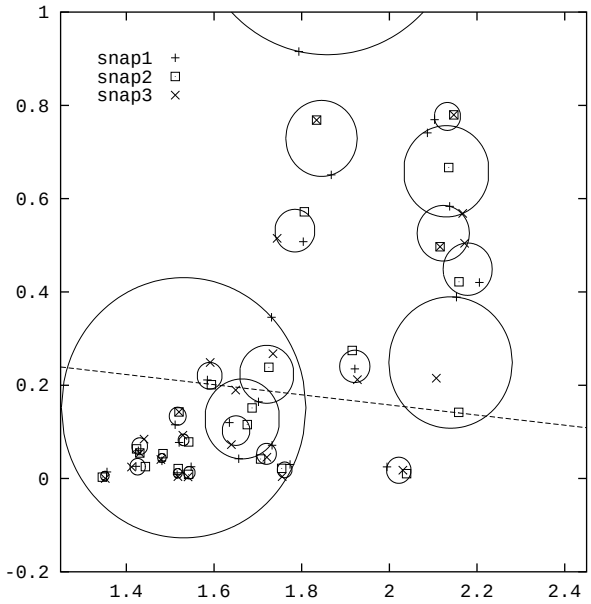

Figure 7: Feature variation under repeated sampling.

CONCLUSIONS

In classical pattern recognition, the Bayes Risk serves as a reference – a limit of excellence that cannot be surpassed. In this paper, we have shown that by relaxing the assumption that the input be sampled only once, a classification system can be built that beats the Bayes error bound.

While our approach to improving recognition accuracy makes use of voting, it is fundamentally different from research on combining the outputs of multiple classifiers (*e.g.*, [2, 4]). Repeated sampling employs just a single classifier, and hence enjoys an attractive property: since there can only be one optimal classifier for a given set of distributions $p(C_i)$ and $p(\hat{\mathbf{x}}|C_i)$, there is no need to "compromise" by incorporating less-than-optimal recognizers in the voting process.

Finally, in this paper we have treated the basic classifier as a "black box" (*e.g.*, in Figure 2). This has the advantage of generality. However, when we know the structure of the feature vector used as input, there is another, straightforward way to apply repeated sampling and "voting" *prior*

to the classification step. If, for example, the observation noise is additive with zero mean, *i.e.*, $\hat{\mathbf{x}} = \mathbf{x} + n$, we can build a system that simply takes the average $\bar{\mathbf{x}}$ over a set of successive measurements $\{\hat{\mathbf{x}}_i\}$. and use the average $\bar{\mathbf{x}}$ as the input to the classifier. Averaging smoothes out the noise in the observation process; it is easily shown that $P(\lim_{m\to\infty} \bar{\mathbf{x}} = \mathbf{x}) = 1$. We are now beginning to examine the tradeoffs between these two approaches to repeated sampling.

References

[1] C. K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6:247–254, December 1957.

[2] V. P. Concepcion and D. P. D'Amoto. Synchronous tracking of outputs from multiple OCR systems. In *SPIE Character Recognition Technologies*, volume 1906, pages 218–228, San Jose, CA, February 1993.

[3] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[4] T. K. Ho, J. J. Hull, and S. N. Srihari. On multiple classifier systems for pattern recognition. In *Proceedings of the 11th International Conference on Pattern Recognition*, pages 84–87, the Netherlands, September 1992.

[5] M. K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, IT-8:179–187, 1962.

[6] D. Lopresti and J. Zhou. Using consensus sequence voting to correct OCR errors. Technical Report 96-94, MITL, April 1994. To be presented at the *IAPR Workshop on Document Analysis*, Kaiserslautern, Germany, October 1994.

[7] J. Zhou and D. Lopresti. Transcending the Bayes Limit through repeated sampling. Technical Report 117-94, MITL, August 1994.