6 Conclusions

We have investigated a new methodology for studying recognition problems — first we construct a realistic generative model, then we empirically study the resulting distributions to estimate tradeoffs between computational requirements and the probability of error. Our experiments show that such simulations can provide a rich source of statistics to guide designers of classifiers, providing worst-case bounds on the computational resources required to achieve specific error and reject rates.

There are many open problems: estimators that can yield tighter bounds should be sought; experiments on more classes, more fonts, and larger images are needed; the defect model must be more carefully validated; and a sampling procedure providing a variance structure that facilitates statistical extrapolation would be helpful.

In this methodology, domain knowledge is expressed in the image defect model, and thus is effectively separated from the methodology for designing classifiers. This offers an escape from the severe practical difficulties that result from the lack of real training data sets of adequate size. We have shown in this paper a successful application of this methodology to a challenging character recognition problem.

Acknowledgements

We are thankful to Mark Hansen, David Ittner, Arnold Neumaier, Daryl Pregibon, and Margaret Wright for their helpful comments.

References

[1] H.S. Baird, R. Fossey, A 100-Font Classifier, *Proceedings of the first International Conference on Document Analysis and Recognition*, St.-Malo, France, September 20–October 2, 1991, pp. 332-340.

[2] H.S. Baird, Document Image Defect Models, in H.S. Baird, H. Bunke, K. Yamamoto (Eds.), *Structured Document Image Analysis*, Springer–Verlag, 1992, pp. 546-556.

[3] H.S. Baird, Document Image Defect Models and Their Uses, *Proceedings of the Second International Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan, October 20–22, 1993, pp. 62–67.

[4] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, Addison-Wesley, New York, 1973.

[5] D.J. Hand, Recent Advances in Error Rate Estimation, *Pattern Recognition Letters*, **4**, 1986, pp. 335-346.

[6] T.K. Ho, H.S. Baird, Asymptotic Accuracy of Two–Class Discrimination, *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, April 11-13, 1994, pp. 275-288.

[7] S.V. Rice, J. Kanai, T.A. Nartker, An Evaluation of OCR Accuracy, in *Information Science Research Institute, 1993 Annual Research Report*, University of Nevada, Las Vegas, 1993, pp. 9–20.

[8] G.T. Toussaint, Bibliography on Estimation of Misclassification, *IEEE Transaction on Information Theory*, **IT-20**, 4, July 1974, pp. 472-479.

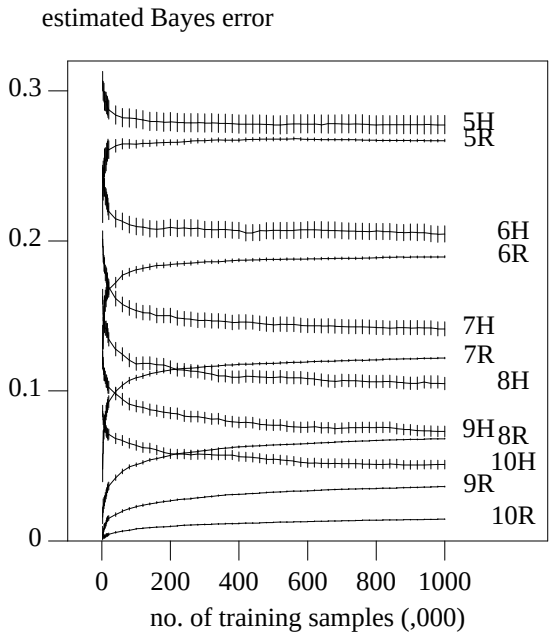

Figure 6: R and H estimates of Bayes error for different text sizes, shown with 95% confidence intervals.

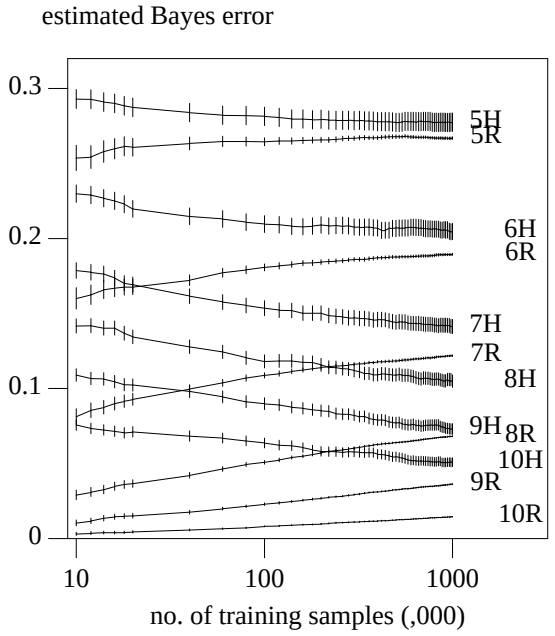

Figure 7: R and H estimates of Bayes error for different text sizes, shown on log-scale of n.

lems are less difficult than they have been imagined. For instance, the best achievable accuracy for 5 point images (mostly smaller than 5×5 pixels), under our defect model, is between 71.6% to 73.4% (26.6% to 28.4% error), which is far better than a random guess. The impact of text size on the error rate is obvious.

The R and H estimates at $n = 1,000,000$ differ by only 1.03% for the 5 point images, but they differ by 3.63% for the 10 point images. This is so even though we enjoy a large ratio of n to the dimension of the feature vector ($1,000,000/81$). Although the dimension of the feature vector is the same for all text sizes, the smaller images have constant zeros in most of the dimensions, and thus the larger images have a higher *effective* feature dimension, and this may help explain the differences in the gaps. Another explanation is that new samples of larger text sizes are less likely to have occurred in the training set, and thus more often have to resort to nearest-neighbor matching.

Achieving sharper estimates may be costly. An extrapolation based on Figure 7 suggests that, for R and H to converge to within 1% on 10 point text, between 10 and 100 million samples may be needed. Of course, more nearly unbiased estimators would help. It seems that extending these experiment to images of higher resolution, and to more classes, may prove challenging.

5 Implications for Classifier Design

The classifier being studied is a brute-force design: it simply stores all training samples and record with each the frequency of occurrence of each class. A new image is compared to all the stored samples. If no match is found then the case can be rejected or a nearest-neighbor is located instead. The time and space requirements of such a classifier are great, perhaps the greatest of all classification methods.

Given our essentially unlimited source of training samples, and assuming we are given sufficient memory, we can make the error rate of this classifier approach the Bayes risk, and the reject rate approach zero (as indicated by the declining fraction of unseen images in Figure 5).

Figures 6 and 7 and their extrapolations show the range of accuracy achievable with a given number of samples. For example, in our problem, the estimated Bayes error for the 10 point images is between 1.43% and 5.39% with 1,000,000 training samples. If we reject when no exact match is found, then from Figure 5 we know that a new sample of 'c' has a 37% chance of being rejected, and for 'e's it is 48%. To achieve this accuracy one would need to store 509,589 unique training samples (Figure 3) and their corresponding class decisions, which takes $509,589 \times (81(\text{the image}) + 1(\text{the class decision}))$ bits = 5.2 Mbytes.

Table 1: Frequency of occurrence of identical images in both classes (measured with 500,000 samples of each class).

Text size	5 pt	6 pt	7 pt	8 pt	9 pt	10 pt
1) Number of shared images in two classes	2,825	8,127	14,113	17,288	14,446	7,917
2) % all distinct images	29.6%	24.3%	15.1%	8.2%	4.1%	1.6%
3) % all distinct images of 'c's	47.6%	42.8%	30.4%	18.0%	8.9%	3.4%
4) % all distinct images of 'e's	43.9%	36.0%	23.1%	13.2%	6.9%	2.8%
5) Corresponding number of samples	987,314	935,117	743,299	478,838	236,578	90,882
6) % all samples	98.7%	93.5%	74.3%	47.9%	23.7%	9.1%
7) % shared in all samples of 'c's	99.0%	95.3%	81.8%	57.3%	32.9%	12.7%
8) % shared in all samples of 'e's	98.5%	91.8%	66.9%	38.5%	14.4%	5.5%

identical to a sample of 'c's. This could be caused by the fact that there are more shape variations in 'e's (a larger number of distinct images) than in 'c's, and by the characteristics of our defect model.

Notice that although the distributions overlap heavily, the frequency of occurrence of the images shared by both classes could differ significantly between the two classes. For instance, at 5 point, only 475 (17%) of the 2825 images shared by the two classes occur with equal probability in both classes. This suggests that, by using the frequency information, classification of the confusable images can be more accurate than random guesses. In the next section, we attempt to estimate the minimum probability of error for these samples when frequency is used.

4 Empirical Estimate of Bayes Error

The Bayes probability of error is the minimum probability of error achieved when the Bayes decision rule is used. For discrimination between two classes c_1 and c_2, the Bayes decision rule for a sample x is to

$$decide\ c_1\ if\ p(c_1|x) > p(c_2|x);\ otherwise\ decide\ c_2.$$

For discrete x, the Bayes error is given by [4]

$$P(e) = \sum_{x \in R_1} p(x, c_2) + \sum_{x \in R_2} p(x, c_1),$$

where $R_1 = \{x | p(c_1|x) \geq p(c_2|x)\}$ and
$R_2 = \{x | p(c_2|x) > p(c_1|x)\}$.

We construct a Bayes classifier using this decision rule, where $p(c_1|x)$, $p(c_2|x)$, $p(c_1, x)$ and $p(c_2, x)$ are estimated by frequency substitution with a training sample set of size n. For samples not included in the training set, the probabilities are estimated by the frequencies at their nearest neighbor (under Hamming distance). As n increases, the number of unseen images approaches zero, and so the error rate of this classifier approaches that of an ideal classifier that possesses perfect knowledge of $p(c_1|x)$ and $p(c_2|x)$, that is, it approaches the Bayes probability of error.

Using our data model it is possible in principle to generate samples indefinitely, and so drive up accuracy until it reaches the Bayes limit. However, in practice we are constrained by finite computational resources; thus the rate of change of the accuracy estimate as a function of n is of practical importance. As for the choice of an estimator, many have been proposed in the literature [8] [5]. Here we choose to apply the resubstitution method and the holdout method. In the resubstitution method (we denote it R), the sample set that is used to estimate the posterior probabilities and to construct the classifier is also used to measure the error; this is biased optimistically. In the holdout method (H), a reserved sample set images (we use 10,000), disjoint from those used in constructing the classifier, is used to measure the error; this is biased pessimistically. As sample size n increases (and the number of features is held constant), it has been conjectured that the difference between these two estimates vanishes in the limit [8]. It has also been suggested that a linear combination of the two estimates gives an essentially unbiased estimate for the true error[5].

Figure 6 shows the estimates and the corresponding 95% confidence intervals obtained for different text sizes, and Figure 7 shows the same data on log scale of n (estimates with $n < 10,000$ are omitted for clarity). For a better approximation of normality, the confidence intervals were computed on the transformed statistic $\log(\hat{p}/(1-\hat{p}))$ (where $\hat{p} = 1 - R$ or $1 - H$) and then back transformed. Since the H estimates were computed using fewer samples, their confidence intervals are larger.

We have seen that even the optimistically biased R estimate of Bayes error is nonzero for all our text sizes. Therefore the goal of building perfect classifiers for them is unrealistic. On the other hand, these prob-

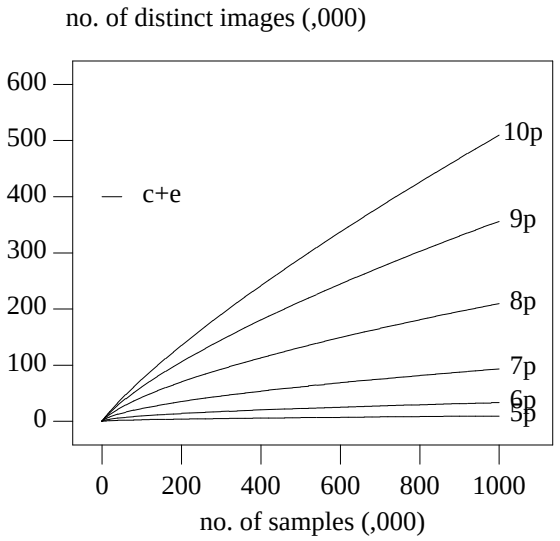

Figure 3: Number of distinct images found in the sample collection.

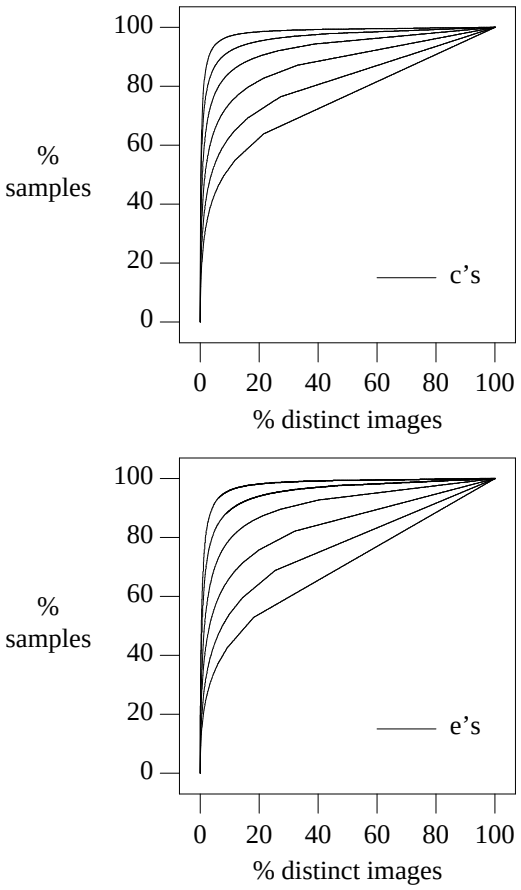

Figure 4: Cumulative frequency of images at 5,6,7,8,9 and 10 pt (from top to bottom).

Assuming that the images are generated randomly, we can ask, at each step, 'what is the probability that the next image will be identical to one already seen?' We estimate this probability by a sliding-window method. Starting with an empty set, we add new samples in batches of fixed size (10,000). Before each new batch is added, we determine how many of the new samples have been seen in any of the earlier batches. The frequency of previously unseen images declines quickly as the sample size grows (Figure 5).

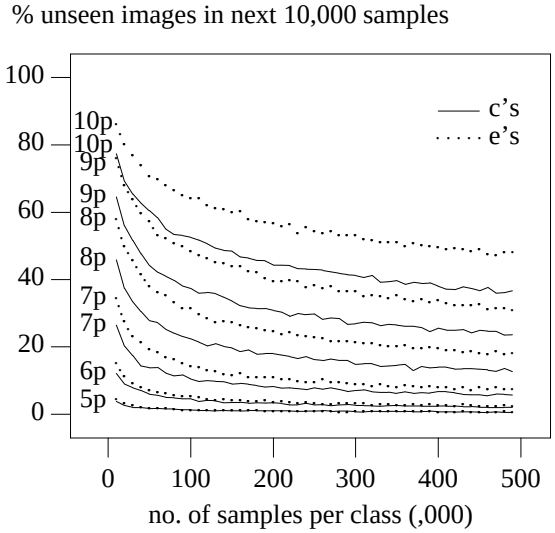

Figure 5: Changes in the fraction of unseen images in each additional 10,000 images.

To assess the difficulty of discrimination it is important to measure the overlap of the class-conditional distributions. One measure of this is the frequency of occurrence of images shared by both classes. For these images, if frequency information is not used, the class can be decided only at random. Table 1 summarizes the overlap statistics for our entire collection of samples.

Table 1 shows that, at 5 point, the distributions overlap heavily. Identical pairs can be found between 99.0% of the 'c's and 98.5% of the 'e's (lines 7 and 8), and only 1.3% of the samples can be uniquely classified. On the other hand, at 10 point, only 9.1% of all samples (line 6) are potentially confusable. This gives a measure of the difficulty of the problem, and it is clearly affected by the text size.

Another interesting observation from Table 1 is the asymmetry between 'c's and 'e's. For all text sizes, larger fractions of 'c's are confusable with 'e's than vice versa. In other words, 'e's are less likely to be

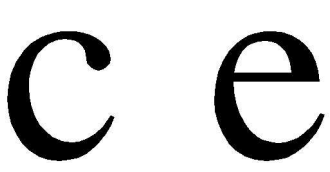

Figure 1: Ideal images of 'c' and 'e' in the Adobe Times Roman typeface.

this model, scored 99.7% accuracy on real images of English books [1]. More thorough validation of such models is currently an active research topic.

Text size has a marked effect on OCR accuracy [6]. To study the difficulty of recognition as a function of text size, we created 1,000,000 images (500,000 for each class) at each of six text sizes (5, 6, 7, 8, 9 and 10 point). The images were generated at a spatial sampling rate of 100 pixels/inch (ppi), and thus are similar to those occurring in challenging real–world OCR problems, such as those arising in low–resolution FAXes (100×200 ppi). Figure 2 shows some of these sample images.

Among images at this resolution and within this range of text sizes, the vast majority (99.98%) fit within a grid of 9×9 pixels and thus can be represented by an 81-component binary vector (we center images in the grid). The empirical distribution of samples within this 81-dimensional space is the principal object of our study. We are interested in those characteristics of the distributions that affect both the accuracy and time/space demands of a classifier.

3 Characteristics of Sample Distributions

The first characteristic we investigated was the frequency of occurrence of an image in each class. In an 81-dimensional binary space, there are at most $2^{81} \approx 2.4 \times 10^{24}$ distinct points each corresponding to an image. The large volume of the space has led to a belief that classifying by exact match is infeasible due to the hopelessly large number of prototypes needed. Yet not every one of these images has an equal probability of occurrence in each class. If the distributions have a strong central tendency, a limited number of prototypes might be sufficient to represent a large fraction of all images that are likely to occur, so that

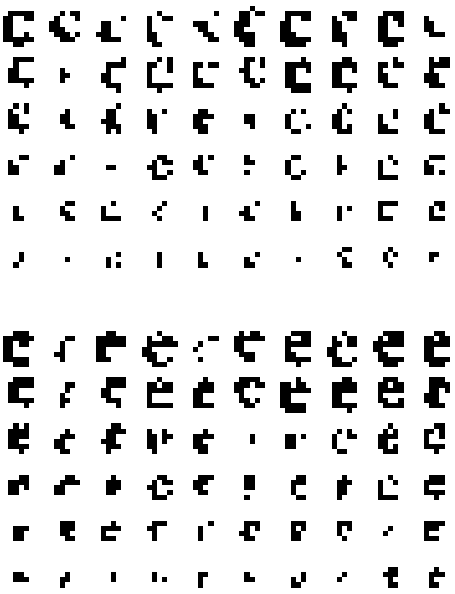

Figure 2: Sample images of c's (upper group) and e's (lower group) at 100 ppi and in 10, 9, 8, 7, 6, and 5 point (from top to bottom in each group).

at least a partial solution to the recognition problem can be obtained within practical limits.

We summarize our findings as follows. The distribution of 5 point images is the most compressed — out of the 500,000 samples of 'c's, only 5,939 distinct images were seen; and of the 500,000 samples of 'e's, only 6,431. On average the ratios of overdetermination are 84 and 78 respectively. Even for the 10 point images, which is the most dispersed, a significant number of repetitions are observed. Out of the 500,000 samples of each class, there are 230,054 distinct 'c's and 287,452 distinct 'e's, and the ratio of overdetermination is 2.17 and 1.74 respectively. Figure 3 shows the increase in the number of distinct images as a function of the sample set size.

The occurrence of each distinct image in the sample set is highly nonuniform. Figure 4 shows the cumulative frequency of the sample distributions, where distinct images are ordered by their frequency of occurrence. For instance, 90% of the 5 point samples are identical to one of less than 4% of the distinct images at that size. This central tendency is stronger for the 'c's than for the 'e's. At 10 point, 60% of the 'c's are included in 18% of the distinct images, but the same number of the 'e's are distributed over 30% of the distinct images.

Estimating the Intrinsic Difficulty of A Recognition Problem

Tin Kam Ho, Henry S. Baird

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974, USA

Abstract

We describe an experiment in estimating the Bayes error of a concrete image classification problem: a difficult, practically important, two–class character recognition problem. The Bayes error gives the "intrinsic difficulty" of the problem since it is the minimum error achievable by any classification method. Since for many realistically complex problems, deriving this analytically appears to be hopeless, we approach the task empirically. We proceed first by expressing the problem precisely in terms of ideal prototype images and an image defect model, and then by carrying out the estimation on pseudorandomly simulated data. Arriving at sharp estimates seems inevitably to require both large sample sizes — in our trial, over a million images — and careful statistical extrapolation. The study of the data reveals many interesting statistics, which allow the prediction of the worst–case time/space requirements for any given classifier performance, expressed as a combination of error and reject rates.

1 Introduction

The machine vision scientific and engineering communities suffer, we believe, from a scarcity of performance guarantees. By this we mean that potential users of machine vision technology can rarely expect definite answers, without first carrying out expensive custom engineering trials, to questions such as the following.

- Can a machine be built for my problem which will achieve accuracy X?
- If so, what are its runtime/memory requirements?
- What is the best accuracy possible in practice on my problem?
- What is the best accuracy possible in principle on my problem?

The intractability of these questions results from many causes, including the difficulty of unambiguously specifying vision problems, the immature state of methods for estimating the difficulty of problems, and the internal complexity of machine vision systems.

In this paper, we describe first steps towards a methodology for answering such questions, and its application to a concrete, realistically complex, practically important problem in character recognition. We hope that this work lays the foundation for a method of automatically constructing classifiers that are guaranteed to achieve any user–specified accuracy, limited only by the problem's intrinsic difficulty and, of course, the computational resources available.

2 Data Generation

We consider the problem of discriminating the lower–case letters 'c' and 'e' in the Adobe Times–Roman typeface. This problem is of practical interest, difficult but not hopeless, and is not easily resolved by geometric context [6]. A study of commercial OCR machines [7] ranks 'e'→'c' and 'c'→'e' as the 2nd and 14th most common confusions.

We define this problem precisely in terms of ideal image artwork (Figure 1) that is degraded by a quantitative model of the physics of printing and imaging [2] [3]. The model specifies a distribution on a number of distortion parameters; by sampling pseudorandomly from this multi-variate distribution, we generate an indefinitely long sequence of distorted images. This defect model is designed to produce shape distortions similar to those occurring in real-world document images. A rough validation of the model has been carried out in experiments in which a page reader, using a classifier trained only on synthetic data from