Cross Validation Comparison Of NIST OCR Databases

Patrick J Grother
Image Recognition Group
National Institute of Standards and Technology

Abstract

The quality of reference databases for Optical Character Recognition is vital to the meaningful assessment of classification algorithms. The National Institute of Standards and Technology (NIST) has produced two databases of segmented handprinted characters obtained from socially distinct writer populations. Two approaches to the comparison of the databases are described. The first uses the eigenvalue spectrum of the covariance matrix as an *a priori* measure of the variance intrinsic to the data. The second cross validates the datasets using classification error to quantify the difficulty of OCR.

The eigenvalue spectra from the training partitions of the datasets are generated during the production of the Karhunen Loève (KL) Transforms, the leading components of which are used as prototype features for a classifier. The eigenspectra are used to quantify diversity of the character sets and the Bhattacharrya distance is used to measure class separability.

The digits, upper and lower case letters from the two populations of 500 writers are partitioned into N disjoint sets. The KL transforms of each such set are used for testing, while the remaining $N-1$ sets form the training prototypes for a Probabalistic Neural Network (PNN) neighbour classifier. Recognition error rates and their variances are calculated over the N partitions for both databases independently. This quantifies intra-database diversity. The inter-database results, or "cross" terms, obtained by training and testing on different databases, indicate the generality of the training set.

The results for digits suggest that the second NIST database (used nominally for testing) is significantly harder than the first (training) set; the testing images are 11% more variable. The NIST training data classifies partitions of itself with 1.7% error, and the test set with 6.8% error. Conversely the test set generalizes to both itself and the training data with 3.5% error. This effect has also been reported using non-NIST classifiers.

1 Introduction

In the spring of 1992 the National Institute of Standards and Technology hosted the First Census Optical Character Recognition (COCR) Conference [1]. One aim of the conference was to ascertain the state-of-the-art industrio-academic performance on the recognition of NIST segmented numeric, upper and lower case letters.

Participants in the conference agreed to classify unlabelled images using their proprietary and/or public domain recognition systems and submit their classifications to NIST for scoring. NIST provided two databases to all entrants. The first, termed NIST Special Database 3 (SD3) [2], contained the segmented characters of 2100 writers and the "known" class files. This constituted an optional training set. The second database, termed NIST Test Data 1 (TD1) contained unlabelled characters from 500 writers, and constituted the test materials.

One result of the Conference was that those recognition systems trained solely on the NIST SD3 database generally displayed inferior TD1 recognition to those trained on a superset of this data, i.e. one including SD3 as a subset or other, possibly proprietary, datasets. The notion that the SD3 was "clean" or "constrained" relative to the TD1 dataset was suggested by the writer profiles: SD3 was obtained from motivated permanent census field personnel whereas TD1 was obtained from variously motivated, more diverse and cosmopolitan high school students. An example is that the European crossed seven is far more abundant in TD1 than SD3.

This study was initiated to formally investigate the relative differences between the databases. The intent was to obtain some classifier independent measures of the relative database difficulty - to obtain results that pertain to the properties of the data, and not the particular recognition algorithm. Cross validation [3] [4] has long been used as a method of obtaining more "mileage" from a data set; by partitioning the data into disjoint subsets, one for parameter estimation (ie training) and the other for performance measurement (ie testing), more robust estimates of performance statistics are available. An alternative approach avoids classification altogther, instead favouring statistical description of the patterns sets.

2 Theory

Whereas Moody [5] expressed cross validation in terms of the mapping error associated between inputs and targets to a multilayer perceptron, the concept of cross validation is in no way restricted to neural network classifiers or function approximators. Cross validation is a method for accumulation of a statistic which, in the case reported here, is the classification error as obtained using a nearest neighbour classifier.

A problem associated with Multi Layer Perceptron (MLP) networks is that patterns (for example, crossed sevens) present in a training set only in small numbers are only weakly represented by the estimated weights, such that generalization is poor. Non-parametric methods do not model the training data and are thus not usually prone to this problem. Such a method is the ubiquitous K-nearest neighbour algorithm [6]. The distances of an unknown pattern to elements of a prototype set are calculated using a suitable, often euclidean, metric. Voting between the classes of the K closest implies the class of the unknown. Numerous extensions to the scheme have been used effectively including an elaboration, termed "Probabilistic Neural Network" (PNN), due to Donald Spect [7], in which all prototypes are included in a Gaussian distance weighted metric thereby emulating the Parzen density approach. The advantage of the method is that an *a posteriori* probability is attached to each possible class; the unknown is classed as that with the highest probability. NIST has used nearest neighbour classifiers that significantly outperform the MLP networks given identical features.

3 Classification

The first stage of classification used the Karhunen Loève expansion of the images as a reduced dimensionality optimally compact representation. The use of such features in OCR has been described in, for example [8] [9]. The handwritten binary characters are isolated and represented as the ± 1 elements of a column vector by some consistent ordering of the square image. The mean vector of P such images is subtracted from each and an ensemble matrix, $\mathbf{U}$ is formed with these P vectors as its columns. The symmetric covariance matrix, $\mathbf{R}$, gives the mean over all images in the ensemble, of all the interpixel correlations. As such it statistically describes how handwritten character images vary.

$$\mathbf{R} = \frac{1}{P}\mathbf{U}\mathbf{U}^T \tag{1}$$

The covariance matrix $\mathbf{R}$ has eigenvectors as the columns of Ψ defined as:

$$\mathbf{R}\Psi = \Psi\Lambda \tag{2}$$

where the only non zero elements of Λ are the eigenvalues on its diagonal. The eigenvectors are the directions of maximum variance in the image space and form a complete orthonormal set termed the principal axes of a hyperellipse in that space. The eigenvalues define the statistical "length" of these axes: thus the first column of Ψ corresponding to the largest eigenvalue is the major axis. The eigensolution of the covariance matrix provides an ordered variance expansion of the image ensemble. The latter eigenvectors, describing very little variance in the images, are discarded thus affording reduced dimensionality. *Any* image vectors as a column of a new matrix $\mathbf{U}$ is a linear superposition of the basis vectors:

$$\mathbf{U} = \Psi\mathbf{V} \tag{3}$$

where the inversion of this formula, $\mathbf{V}$, defines the Karhunen Loève Transform, the elements of which are the components of the image vector onto the principal axes:

$$\mathbf{V} = \Psi^T\mathbf{U} \tag{4}$$

These feature vectors are classified using the PNN nearest neighbour technique[7]. Although many variations have been described the NIST_4 implementation is as follows. The square euclidean distance of an unknown pattern, $\mathbf{v}$, to the i^{th} prototype of the training set, t_i, is

$$d_i^2 = \sum_{j=1}^{N}(v_j - t_{ij})^2 \tag{5}$$

The distances $d_i \forall i$ are expressed as a function of the standard deviations of normal distributions centered on each of the prototypes. A Gaussian is applied as a kernel weighting function.

$$g_i = e^{-d_i^2/2\sigma^2} \tag{6}$$

The weighted distances are then accumulated by-class over the K classes, to which the prototypes belong.

$$p_k = \sum_{i}^{P} g_i\delta_{ki} \tag{7}$$

where δ_{ik} is unity if the i^{th} prototype is of class k and zero otherwise. Interestingly this vector may be normalized to give true *a posteriori* probabilities by dividing by $\sum_k p_k$. For optimal classification it is necessary to survey over the Gaussian width σ: for digits the best value was taken as 3.0 whereas for uppers and lowers a value of 4.0 was adopted. Note that as $\sigma \to \infty$, $g_i \to 1$ and classifications defaults to that class with the highest *a priori* probability obtained from the row sums of δ.

Rather than use classifiability as a measure of database homogeneity it is possible to obtain *a priori* measures. Consider the databases as image ensembles for which the Karhunen Loève Transform (KLT) is defined [10]. The variances of the transform coefficients are the eigenvalues. Since the eigenvectors,

the basis of the KLT, form a complete orthonormal set, *any* image (including those of the ensemble from which the covariance matrix is calculated) is exactly a linear superposition of those bases. If the eigenvalue spectrum is relatively flat then the variance in an image ensemble is distributed over many eigenvectors and more are needed for an adequate representation, as for instance in achieving a low reconstruction mean square error level. Note that the total image variance is related to the scatter of the data, S, defined as

$$S = E\{ \|\mathbf{u}_i - \mathbf{u}_j\| \} \tag{8}$$

where the expectation, $E\{\cdot\}$, of the underlying distribution is replaced by the sample mean whence

$$S = \frac{1}{P^2} \sum_{i=1}^{P} \sum_{j=1}^{P} (\mathbf{u}_i - \mathbf{u}_j)^T (\mathbf{u}_i - \mathbf{u}_j) \tag{9}$$

$$S = \frac{1}{P^2} \sum_{i=1}^{P} \sum_{j=1}^{P} (\mathbf{u}_i^T \mathbf{u}_i + \mathbf{u}_j^T \mathbf{u}_j) - \frac{1}{P^2} \sum_{i=1}^{P} \sum_{j=1}^{P} (\mathbf{u}_i^T \mathbf{u}_j + \mathbf{u}_j^T \mathbf{u}_i) \tag{10}$$

Given that the $\mathbf{u}$ are mutually independent and from a single distribution the double sums are replaced thus

$$S = \frac{2}{P} \sum_{i=1}^{P} \mathbf{u}_i^T \mathbf{u}_i - \frac{2}{P} \sum_{i=1}^{P} \mathbf{u}_i^T \frac{1}{P} \sum_{j=1}^{P} \mathbf{u}_j \tag{11}$$

The latter sums are the mean vectors defined prior to equation 1 to be zero. The first sum of inner products is the also the trace of the sum of the outer products which is identically the covariance matrix.

$$S = 2\, trace\ \mathbf{U}\mathbf{U}^T \tag{12}$$

The diagonal elements of the covariance matrix are the variances of the image pixels. Given that the total variance is conserved under unitary transformation:

$$\sum_i \mathbf{R}_{ii} \equiv trace\ \mathbf{R} = trace\ \Lambda \tag{13}$$

It is found that the scatter statistic is twice the sum of the eigenvalues. Further expressing the eigenvalues as a percentage of their total yields the percentage of the ensemble variance that is represented by a subset of N eigenvectors. For comparison of the two databases the difference in the percentages as a function of N is considered. If an eigenspectrum is wide, then the percentage variance described by the N leading eigenvectors will be small. If the cross validation percentage classification error is also low the information discarded by using an incomplete KL transform is irrelevant even though there is much of it. Alternatively, if the eigenspectrum is narrow, with much of the variance captured, then a low recognition rate implies that the discarded transform coefficients are valuable. This latter sensitivity to the high order KL-transform is undesirable since the motivation for feature extraction is reduced dimensionality.

Consider the two class problem. Throughout assume that the a priori probability of each class is identical and therefore 0.5. Let the conditional density functions for the two classes be $p_1(\mathbf{u})$ and $p_2(\mathbf{u})$ such that the *a posteriori* probabilities of an example being of class k is

$$q_k(\mathbf{u}) = \frac{p_k(\mathbf{u})}{p(\mathbf{u})} \tag{14}$$

where $p(\mathbf{u})$ is the mixture density function $p_1(\mathbf{u}) + p_2(\mathbf{u})$. The decison rule for classification of an unknown vector $\mathbf{u}$ then becomes simply to choose class 1 if $p_1(\mathbf{u}) > p_2(\mathbf{u})$ and vice versa. The usual problem of not possessing the density functions is obvious. This rule will still not generally give zero error - given a vector $\mathbf{u}$ consider the *conditional error* to be

$$r(\mathbf{u}) = min(\ q_1(\mathbf{u})\ ,\ q_2(\mathbf{u})\ ) \tag{15}$$

and the total (Bayes) error to be the expectation of $r(\mathbf{u})$ thus

$$\epsilon = \int r(\mathbf{u})p(\mathbf{u})\,d\mathbf{u} \ = \ \int min(\ p_1(\mathbf{u})\ ,\ p_2(\mathbf{u})\ )\,d\mathbf{u} \tag{16}$$

$$\epsilon = \int_{L_2} p_1(\mathbf{u})\,d\mathbf{u} \ + \ \int_{L_1} p_2(\mathbf{u})\,d\mathbf{u} \tag{17}$$

where the volumes L_k correspond to classification of $\mathbf{u}$ as class k. This is still of little utility given the inaccessability of the density functions. However it is possible to compute an upper bound on ϵ by using the lemma that $min(a,b) \leq a^s b^{1-s}$ for $0 \leq s \leq 1$. The upper bound on the error then reduces to

$$\epsilon_u = \frac{1}{2}\int p_1^s(\mathbf{u})\,p_2^{1-s}(\mathbf{u})\,d\mathbf{u} \ = \ \frac{1}{2}e^{-c(s)} \ \geq \ \epsilon \tag{18}$$

and for a normal distribution the $c(s)$ can be obtained eventually as (see [11])

$$c(s) = \frac{s(1-s)}{2}\,(\mu_2 - \mu_1)^T\,\mathbf{R}_{12}^{-1}(s)\,(\mu_2 - \mu_1) \ + \ \frac{1}{2}\ln\,\frac{det\,\mathbf{R}_{12}(s)}{det\,\mathbf{R}_1^s\ det\,\mathbf{R}_2^{1-s}} \tag{19}$$

where $\mathbf{R}_{12}(s) = s\mathbf{R}_1 + (1-s)\mathbf{R}_2$, and μ_k and $\mathbf{R}_k$ are the mean vectors and covariance matrices of class k. The left hand side is termed the Chernoff distance and is parameterized by s. It reaches a maximum at $s = 0.5$ only when $\mathbf{R}_1 = \mathbf{R}_2$ in which case the term Bhattacharyya distance is used. These distance measures are useful in defining class separability and are reasonable even when the data is not normal although strictly tests for normality should be applied (The *Kolmogorov-Smirnov* test for example, in [12]).

The first term of equation 19 measures the distance between the class means normalized by the mixture covariance $\mathbf{R}_{12}$. The second term quantifies the distance due to differences in the covariances. The values of ϵ_u are given in table 2 for the ten classes of the two NIST data sets. Note that the matrix is not symmetric since the upper and lower triangles refer to different databases.

4 Cross Validation Results

This study generated a *Validation Comparison Matrix*. The matrix has rank two and dimension equal to the number of databases in the comparison which in this case is also two. The row and column indices of the matrix denote, respectively, the databases used for training and testing. The absolute classification error values in the matrix are irrelevant since the entries were all produced using identical classifiers none

of which were particularly optimized. The interesting features are the relative percentages discussed below.

The on-digaonal terms, c_{ii}, indicate the mean result for standard v-fold cross validation of the i^{th} database. The off-diagonal elements, c_{ij} $i \neq j$, result from cross-cross validation. The first u partitions of database i are used as training sets for the v-fold cross validation of the j^{th} database. In the case of v-fold partitioning of the training set there will be uv results the mean of which is c_{ij}.

All elements have an attached sample standard deviation.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{20}$$

If an homogenous training set is large enough then recognition error and this deviation will approach zero. The standard deviation is also affected by the data set redundancy. For instance, consider a database to which a copy of itself is appended, and which is classified with, for example, a single nearest neighbor algorithm. Perfect recognition could then be achieved if, as in the cross validation scheme used here, the partitions are contiguous blocks from the dataset.

For comparing the means of two different databases the standard error is used. It is accessible by dividing the standard deviations by a further $\sqrt{N}$. The discussion of the comparisons of the means and the variances is aided by invoking the results of Student's "t-test" and the "F-test" (see for example [13]) which utilize this quantity and not the σ values of equation 20.

They are used to assess whether two distributions have the same mean and the same variances. The entire corpus of human hand-printed characters may be considered as one distribution of which SD3 and TD1 are subsets, but for this study the two sets are extracted from different distributions, namely the characters of the two social writer groups outlined in the introduction. The t-test quantifies the difference in two means as a multiple of their mutual root mean square standard errors.

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \tag{21}$$

Attached to it is a significance, $0 \leq p \leq 1$, giving the probability that $|t|$ could be at least this large by chance. That is if p takes on a "small" value then the distributions have "markedly" different means. Similarly the F-test quantifies two variances as a ratio taken to be greater than 1 (i.e. either σ_1^2/σ_2^2 or its reciprocal). The value of F directly indicates differing variances. The attached significance, p is again a probability. Small values indicate significantly different variances.

The statistics are derived from the two samples obtained by testing the 10 partitions of SD3 and TD1 data using one or other training set. In all cases, digits, upper-case and lower-case letters, the calculated value of the t is found with very low significance implying the mean differences are not at all spurious. However in no case does the attached probability for the F-test indicate that the variances are significantly different.

4.1 Digits

The handwritten digits of the first 500 writers of NIST Special Database 3 (SD3) were partitioned into blocks from 50 writers. The number of characters in these ten sets were not identical but varied by only 0.2%. The number of SD3 digits totalled 53449. The 500 writers of NIST Test Data 1 (TD1) were similarly partitioned. The number of TD1 digits totalled 58646. The pure 10 fold cross validations for

Correct % $\pm$ σ	Test SD3 50 writers	Test TD1 50 writers	
Train SD3 450 writers	1.7% $\pm$ 0.3	6.8% $\pm$ 0.4	t = 28.5 p = 0.0 F = 1.5 p = 0.3
Train TD1 450 writers	3.5% $\pm$ 0.3	3.8% $\pm$ 0.5	t = 1.4 p = 0.2 F = 2.1 p = 0.3

Table 1: Inter and Intra database Cross Validation Recognition Errors for Digits.

SD3 and TD1 were obtained using the characters of 90% of the writers as prototypes for the characters of the remaining 10%. The mean incorrect classification percentages are quoted on the diagonal of table 1.

The first partition only ($u = 1$) of SD3, that is a fixed 450 writers, were used as prototypes for the classification of all $v = 10$ sets of characters of TD1, and vice versa. The off-diagonal elements of the validation comparison matrix, so obtained, are given in the table.

The most relevant result from the above table is that, using the classifier as described above, training solely on SD3 implies a 5% loss when classifying TD1. This is effectively NIST's experience with its NIST_0 and NIST_1 conference entered systems.

The on-diagonal elements of the cross validation matrix show that SD3 is a less *diverse* digit set than TD1. That is the test partitions of SD3 are more like their training sets, in the nearest neighbour sense, than is the case with TD1. Greater on-diagonal terms indicate a higher intrinsic diversity for that database. If we relate the low TD1 classification to the width of the eigenvalue spectrum or the volume of the eigenspace it is apparent that TD1 would benefit from the use of a less incomplete KL transform as input to the classifier.

Figure 1 shows the eigenspectra of the SD3 and TD1 characters. Note in particular the total variances for the 1024 pixel digits are 576 (SD3) and 637 (TD1) indicating that TD1 is absolutely more diverse (larger scatter). Approximately 6.6% more of the variance of SD3 is described by 48 KL coefficients (as used by the classifier) than is the case for TD1.

The off-diagonal terms show that SD3 as prototypes for TD1 is markedly inferior to TD1 as a training set for SD3. The implication is that TD1 is a superset of the SD3 set, i.e. TD1 contains sufficiently distributed prototypes to classify SD3 - whereas TD1 contains exemplars that are not "closely" present in SD3. That TD1 classifies itself and SD3 equally (to within one standard deviation) implies that TD1 is a more *general* dataset.

Table 2 shows the Bhattacharyya upper error bounds for all the digits of SD3 and TD1. A full Chernoff table is expensive to compute so only the $s = 0.5$ value is given which is generally an overestimate of

		TD1									
		0	1	2	3	4	5	6	7	8	9
SD3	0	60.06	0.00	1.53	0.74	0.37	2.06	2.08	0.27	0.52	0.23
	1	0.00	56.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	1.85	0.00	51.74	2.94	1.96	3.23	2.53	3.79	6.71	2.21
	3	0.99	0.00	7.79	57.99	0.96	4.05	0.95	2.75	7.80	3.63
	4	0.59	0.00	2.24	1.00	50.13	3.49	1.94	5.14	3.07	16.30
	5	2.60	0.00	2.96	5.77	3.63	63.12	6.42	1.92	8.55	3.02
	6	0.91	0.00	1.74	0.33	0.81	1.61	44.89	0.10	0.63	0.06
	7	0.21	0.00	3.38	1.26	1.99	1.30	0.08	55.93	1.44	13.54
	8	0.13	0.00	7.59	4.38	2.40	4.72	2.05	2.95	74.93	3.82
	9	0.18	0.00	1.72	1.48	8.36	2.92	0.28	12.88	5.30	71.70

Table 2: Bhattacharyya bounds for digits of NIST Special Database 3 and NIST Test Data 1. The ij^{th} entry gives the upper bound on the classification error between the i^{th} class of SD3 and the j^{th} class of TD1 given all their examples in isolation.

the upper error bound for any given entry. It is found in the case of digits of different class that the first "mean-difference" term of equation 19 is dominant. The exceptions are in the discrimination of any class from class "1", and from digits of the same class obtained from differing databases when the second covariance-difference term is larger. Note the strong overlaps off-diagonal - the "4" - "9", "5" - "8", "3" - "8" and the "2" - "8" possses small Bhattacharyya distance. Uniquely ones are well separated from all else. If the off-digaonal row (or column) elements, reweighted by *a priori* probabilities of 0.1 not 0.5, are summed, then the total of these is 4.42%. This figure and the individual error levels are higher than was typically achieved using real classifiers. This overestimation of the error is due to either the normally distributed model being inappropriate, or the classes are separable by an alternative (nonlinear) classifier, or the values at $s = 0.5$ are well above their minima. The on-diagonal terms compare the examples from two distributions for a given class. If two distributions are identical then a usual distance measure will yield 0 and the mixing error will be 100%. In this case the large on-diagonal values indicate that the data from SD3 and TD1 are at least partly separable and therefore different. The magnitudes of the terms in equation 19 indicate that this is largely due to differences in the covariances.

4.2 Uppers

The handwritten uppers of the first 480[1] writers of NIST Special Database 3 (SD3) were partitioned into blocks from 48 writers. The upper case letters totalled 10790 examples. The 500 writers of NIST Test Data 1 (TD1), similarly partitioned, yielded 11941 characters. As in the case of digits there is a 5% difference between the classification of SD3 on itself and on TD1. Again TD1 is more diverse in classification of itself than is the case with SD3. The total variances are 734 (SD3) and 650 (TD1) indicating that SD3 is absolutely more diverse. With classification using 96 KL coefficients the percentage variance captured for SD3 was 4.8% *less* than that for TD1. This is reconcilable by considering the separability of the inter-class separability of the SD3 set to be greater while intra-class variation is larger.

The off-diagonal elements, however, are the same indicating that neither set is more general than the other. That the off-diagonal elements are higher than the on-diagonals indicates the databases contain unique subsets that require "specialist knowledge" contained only in that database.

[1]None of the uppercase letters of 20 writers were successfully segmented in the preparation of SD3. See section 5.

Correct % ± σ	Test SD3 48 writers	Test TD1 50 writers	
Train SD3 432 writers	14.2% ± 1.4	19.4% ± 1.4	t = 7.9 p = 0.0 F = 1.0 p = 0.8
Train TD1 450 writers	19.3% ± 1.7	16.5% ± 1.4	t = 3.8 p = 0.0 F = 1.5 p = 0.4

Table 3: Inter and Intra database Cross Validation Recognition Errors for Uppers.

4.3 Lowers

The handwritten lowers of the first 490[2] writers of NIST Special Database 3 (SD3) were partitioned into blocks from 49 writers. The lower case letters totalled 10968. The 500 writers of NIST Test Data 1 (TD1), similarly partitioned, yielded 12000 characters. In particular the total variances are 740 (SD3) and 638 (TD1) indicating that SD3 is absolutely more diverse. With classification using 96 KL coefficients the percentage variance captured for SD3 was 2.3% less than the that for TD1. The cross validation matrix shows that the lower case datasets are equally difficult and yet different - they are insufficiently general to classify the other as well as they classify themselves.

5 Caveats

5.1 Segmentation

This paper is an initial report into the work NIST conducted immediately after the COCR conference. As such it is a provisional investigation of database quality. It is reasonable to conclude that the digits of SD3 are cleaner than those of TD1. However the study is not experimentally flawless - it is not a conclusion that the writers of SD3 em wrote characters neater than those of TD1, only that the characters ultimately included in the database are cleaner. One reason for this is that SD3 and TD1, both obtained from fields of full page forms, were arrived at with different segmenters. From a possible 65000 characters on each 500 form set, final numbers of human checked characters were 53449 (SD3) and 58646 (TD1). The SD3 segmentor, an old version, produced 9% fewer isolated characters than the updated model used for TD1, the principal reason for failure being connected characters. If the characters from SD3 that were not segmented resemble the difficult images that putatively characterize TD1 then the difference between the two databases may not be writer-letter dependent at all, rather it would be a function of the writer-connectivity that different writer groups use.

[2]None of the lowercase letters of 10 writers were successfully segmented in the preparation of SD3. See section 5.

Correct % ± σ	Test SD3 49 writers	Test TD1 50 writers	
Train SD3 441 writers	19.6% ± 1.4	23.5% ± 1.4	t = 5.9 p = 0.0 F = 1.1 p = 0.8
Train TD1 450 writers	25.9% ± 1.8	19.2% ± 1.1	t = 9.6 p = 0.0 F = 2.5 p = 0.1

Table 4: Inter and Intra database Cross Validation Recognition Errors for Lowers.

This problem could be negated by resegmenting and rechecking SD3 using the identical algorithms applied to TD1. A new different database, a superset of SD3, is then obtained, which can then be used in a more controlled comparison with TD1.

The Chernoff integral of 18 is approximated using the normal distribution assumption to obtain an analytic easily computable expression. This assumption is certainly untrue for binary images. The sensitivity of the values in the table to the parameter s was not computed in this study.

5.2 Classifier Dependence

The eigenvalue spectrum describes the information loss suffered when only the leading KL coefficients are used in classification. The classification of incomplete KLT's is peculiar in that variance ordered information is discarded. It is not clear that using a much higher number of coefficients in the digit classification will not equalize the on-diagonal cross validation entries. Even though the nearest neighbour recognition of minority patterns in a higher dimensional (but lower variance) KL space is not possible for parameterized classifiers, the nearest neighbour schemes do better.

Instead of using a "lossy" incomplete feature classifier it is possible to instead use a full description of the image; the complete KL transform. *Variance equalization* may be more reasonable - choose the number of KL features corresponding to either an absolute level of described variance or percentage thereof. Thus in the case of digits, 43 eigenvectors describe 75% of the variance whereas to reach this level with TD1, 70 KL coefficients are required. Alternative features may be used that do not bias information loss. For example image row and column pixel histograms or orthogonal moments are known to be classifiable features for OCR.

6 Conclusions

Given the experimental scheme described, it appears that NIST Test Data 1 is indeed a more diverse and general digit set than NIST Special Database 3. The NIST training digits classified themselves with 5% more accuracy than they classified the test set. Further, the use of NIST Test Data 1 as a training set yields a 3% improvement over Special Database 3 in the classification of that test set. The hypothesis that differing writer populations are responsible for this diversity remains only a possible conclusion. Indeed, the fact that the cross validations for the uppers and lowers yield insignificant differences between the two databases weakens the argument for digits.

References

[1] R. A. Wilkinson, J. Geist, S. Janet, P. J. Grother, C. J. C. Burges, R. Creecy, B. Hammond, J. J. Hull, N. J. Larsen, T. P. Vogl, and C. L. Wilson. The First Optical Character Recognition Systems Conference. Technical Report NISTIR 4912, National Institute of Standards and Technology, August 1992.

[2] C. L. Wilson and M. D. Garris. Handprinted character database. *NIST Internal Report*, 1992.

[3] M. Stone. Cross validatory choice and assessment of statistical procedures. *Journal of the Royal Statistical Society*, B36, 1974.

[4] J. W. Tukey and F Mosteller. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Handbook of Social Psychology*, volume 2. Addison-Wesley, 1968.

[5] J. Moody and J. Utans. Selection of neural net architectures via the prediction risk: Application to corporate bond rating prediction. In *Proceedings of the First Intl. Conference on Artificial Intelligence Applications on Wall St.* IEEE Computer Society Press, 1991.

[6] T. M. Cover P. E. Hart. Nearest neighbour pattern classification. *IEEE Transactions on Information Processing*, IT-13:21–27, 1967.

[7] Donald F. Specht. Probabalistic neural networks. *Neural Networks*, 3(1):109–118, 1990.

[8] P. J. Grother. Karhunen Loève feature extraction for neural handwritten character recognition. In *Proceedings: Applications of Artificial Neural Networks III*. Orlando, SPIE, April 1992.

[9] T. P. Vogl, K. L. Blackwell, S. D. Hyman, G. S. Barbour, and D. L. Alkon. Classification of Japanese kanji using principal component analysis as a preprocessor to an artificial neural network. In *International Joint Conference on Neural Networks I*, pages 233–238. IEEE, 7 1991.

[10] Anil K. Jain. *Fundamentals of Digital Image Processing*, chapter 5.11, pages 163–174. Prentice Hall Inc., international edition, 1989.

[11] K. Fukunaga. *Introduction to Statistical Pattern Recognition*, chapter 3. New York: Academic Press, second edition, 1990.

[12] G. E. Noether. *Elements of Non-Paramatric Statistics*. Wiley, New York, 1967.

[13] S. A. Teukolsky W. H. Press, B. P. Flannery and W. T. Vetterling. *Numerical Recipes*, chapter 13, pages 464–469. Cambridge University Press, 1989.

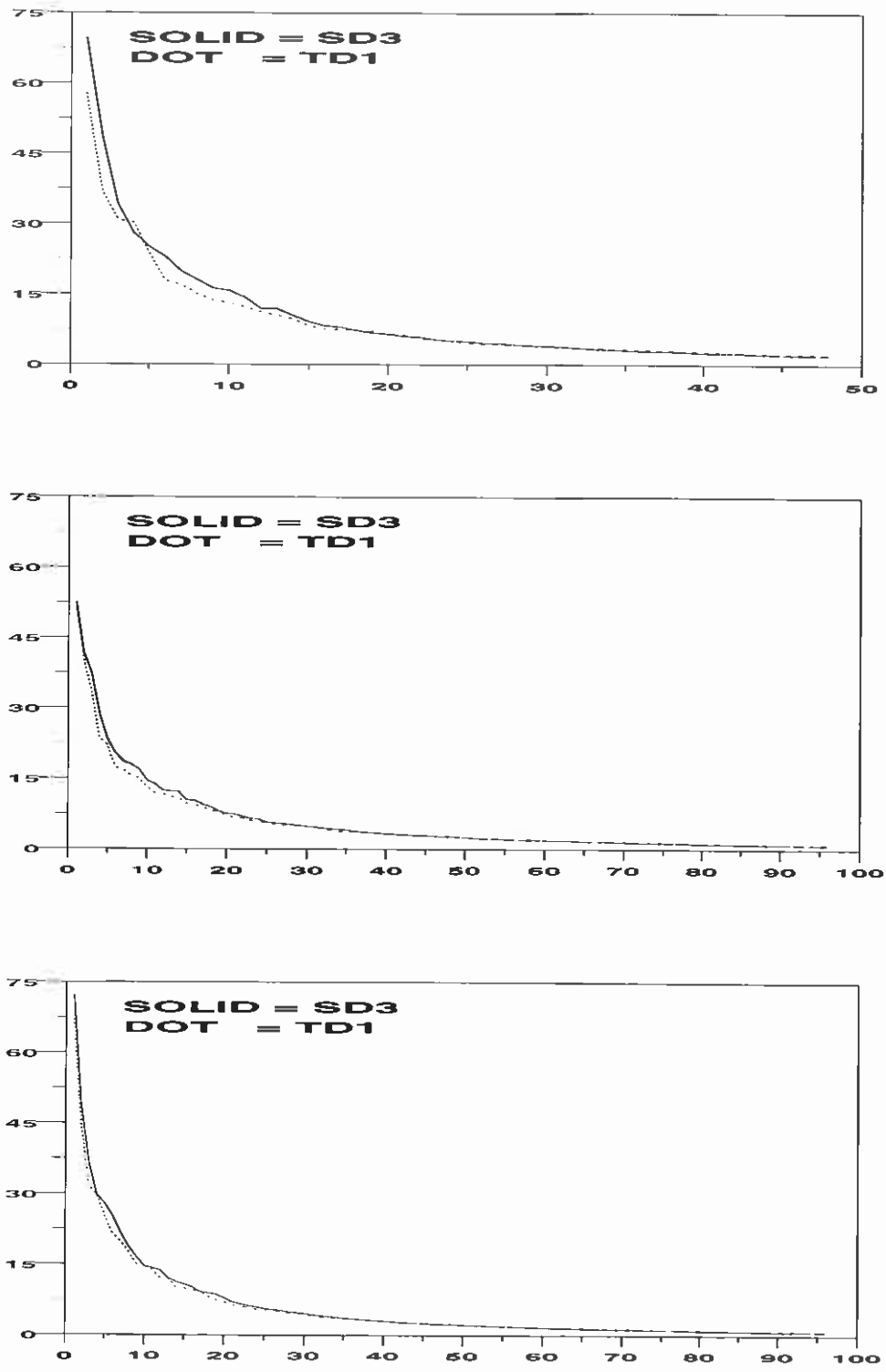

Figure 1: Eigenvalue vs Index for SD3 and TD1. From top digits, uppers and lowers. All writers were used.