NISTIR 4824

Karhunen Loeve Feature Extraction for Neural Handwritten Character Recognition

Patrick J. Grother

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Gaithersburg, MD 20899

NIST

NISTIR 4824

Karhunen Loeve Feature Extraction for Neural Handwritten Character Recognition

Patrick J. Grother

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Gaithersburg, MD 20899

April 1992

U.S. DEPARTMENT OF COMMERCE
Barbara Hackman Franklin, Secretary

TECHNOLOGY ADMINISTRATION
Robert M. White, Under Secretary for Technology

NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
John W. Lyons, Director

Karhunen Loève Feature Extraction For Neural Handwritten Character Recognition

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

Abstract

The optimality of the Karhunen Loève (KL) transform is well known. Since its basis is the eigenvector set of the covariance matrix, a statistical, not functional, representation of the variance in pattern ensembles is generated. By using the KL transform coefficients as a natural feature representation of a character image, the eigenvector set can be regarded as an unsupervised biological feature extractor for a (neural) classifier. The covariance matrix and its eigenvectors are obtained from 76753 handwritten digits. This operation is a unique expense; once the basis set is calculated it forms a linear first layer of a three weight layer feed forward network. The subsequent nonlinear perceptron layers are trained using a scaled conjugate gradient algorithm that typically affords an order of magnitude reduction in computation over the ubiquitous back-propagation algorithm. In conjunction with a massively parallel computer, training is expedited such that tens of initially different random weight sets are trained and evaluated. Increase in training set size (upto 76755 patterns) gives less accurate learning but improved generalization on the fixed disjoint test set. A neural classifier is realized that recognizes 96.1% of 15000 handwritten digits from 944 different writers. This recognition is attributed to the energy compaction optimality of the KL transform.

1 Introduction

1.1 Character Recognition

Optical Character Recognition is normally a multistage process; typically some preprocessing of the image is applied, features are extracted, the result is classified and a rejection decision made. The preprocessing may normalize such attributes as size, field position, rotational orientation and stroke width thus obviating the need for a classifier to be invariant to those transformations. It may also attempt to remove random irrelevant variation from the characters while simultaneously preserving the differences between objects of different classes. Accordingly the machine printed character recognition problem is largely solved because there is little variation in the data[1].

The possible preprocessing operations are numerous. Spacial domain techniques, often used to worthwhile effect, range from rotation or shear, through histogram modification, morpholgical operations, Hadamard Walsh downsampling and neighbourhood averaging, to convolution. Similarly frequency domain methods, such as low pass filtering, can aid noise supression and line connectivity. The transform domain is of particular interest here since the representation typically involves relatively few non zero coefficients. These adequately reconstruct the filtered images since the information content is sufficient to represent the image. Pattern spectra have been widely used in signal and image recognition.

[1] Nevertheless it is always possible to corrupt characters to be worse than any recognizer can cope with; random correlated noise, as introduced by possibly multiple passes through a photocopier, can give sufficient variation in input images to make the machine print problem significant once again.

1.2 Neural Network Approaches

The utility of neural networks in pattern recognition has prompted an effort to model the visual processes that underlie mammalian vision, the intent being to obtain superior artificial recognition networks. Particularly, the receptive fields present in mammalian vision have been proprosed as effective feature extractors for OCR. Foremost among these are the Gabor functions associated by Daugman [2] with the retinal fields of domestic cats. Gabor reconstruction of character images both extracts spatially localized and oriented information and spectrally filters the input data. Alternatively Garris et al [6] used the Gabor transform itself as input to a backpropagation network with notable results.

However all of the above operators are functionally independent of the data; no inherent knowledge of the characters themselves is made use of. A more potent technique, the discrete Karhunen Loève transform (KLT) [7], assumes no model of the human perception mechanism, but more directly references statistically salient information on how handwritten characters are formed. The eigenvectors of the covariance matrix of the character ensemble are taken as a minimal orthogonal basis set, of which any character is a linear superposition. The eigenvectors are the principal statistical components of the variance in the original image space. Their respective eigenvalues indicate the significance of the eigenvector in describing the characters' construction, those with small eigenvalue represent irrelevancies. The motivation for doing this lies not only in the well documented optimality of the KLT [7], but in recent studies [11] showing that evolution of synaptic structures in linear Hebbian neural networks [10] is dynamically governed by the same statistical basis as that of the KLT. Further Vogl et al [18] have described a neural network in which the eigenvectors of Kanji[2] characters evolve during training, and can subsequently be used for classification.

Although eigenvectors appear in some unsupervised network training [15], they are most readily obtained using one of the traditional numerical iterative methods [19]. The eigenvectors can therefore be regarded as a trained weight layer. Martin and Pittman [12] choose to use a two hidden weight layer network with image data as input such that training produces a generally incomplete and non orthogonal basis as the first weight layer. The use of eigenvectors is rather a *prescription* of this feature extraction layer derived as a least mean square fit to the data. This is potentially detrimental to the perceptron as a classifier but it yields pragmatic gains. Perceptron networks are known to exhibit better generalization if the training sets are large. Rather than use raw images as input it is preferable to use greater numbers of precomputed low dimensional KL transforms for training.

In many applications of multilayer perceptrons the classic backpropagation [16] algorithm has been applied with much success. Convergence to error minima during training is notoriously slow and since there is strong evidence that large training sets are important for optimal generalization, it is computationally desirable to use compact representations of images as input to small networks. The starting position in the weight space of the network has a significant effect on the generalization properties of the trained network and, indeed, on the progress of training itself. Expeditious training allows the distribution of generalization perfomance to be estimated over many initial weight sets. NIST has produced serial and parallel Fortran implementations of a new conjugate gradient algorithm [13] [1] that typically affords an order of magnitude reduction in training time over backpropagation[3].

1.3 Experimental Coverage

The experiments reported in this paper investigate the effectiveness of Karhunen Loève transforms as classifiable features for handwritten digit recognition. The issues of interest include:

1. What is the optimal feature length? Generalization on an unseen test database is obtained as a function of the dimensionality of the basis space in which characters are represented;

[2] Actually a subset of the complex Japanese character script.

[3] Send email to James Blue at jlb@azure.cam.nist.gov for NIST Internal Report 4776 and source code.

i.e. the number of classifiable features. Whilst more features more accurately represent a character, too many will describe variation in the characters that is extraneous and redundant. The eigenvalue spectrum of the covariance matrix describes what fraction of the variance is ascribed to a restricted basis subspace. It is known that alphabetic characters yield a wider eigenvalue spectrum due to the increased number of possible strokes. Accordingly Vogl [18] has shown that Kanjii requires still more principal components for its adequate respresentation. The tradeoff between number of features, network size, training ability and generalization is investigated.

2. How big should the training set be? There is some interest in what number of digit exemplars (per class) are required to obtain a statistically robust prototype set, i.e. one which achieves optimal generalization. Network training nonlinearly yields weight sets that are a fixed-size description of the properties of the training data. The nearest neighbour methods (linearly) partition the pattern space by class using, ideally, very large numbers of prototypes. The greater the statistical relevance of a set of prototypes, the less noisy is the error surface defined by those patterns, leading to a better generalization. The drawback of nearest neighbour methods is that they are not adaptive; they are not trained and do not condense representative information from the prototypes and are therefore slow as classifiers. Indeed trained nearest neighbour classifiers are termed neural networks (LVQ and PNN). The nearest neighbour method can be viewed as an untrained form of an adaptive neural multimap pattern recognizer [22] [14] in which the exemplars are not at all aggregated to give some compact representative prototypes. It retains the ability to differentiate between an open top and closed top four whereas a constrained perceptron system is typically, but not necessarily, required to learn to join both subclasses despite their KL transforms being potentially quite different.

2 Karhunen Loève Transformation

2.1 Statistical Representation

Consider that a sample of handwritten characters is available in isolated binary form. These P images $u^{(p)}$ are each of size N by N pixels. The p^{th} character is regarded as a real matrix $\mathbf{u}^{(p)}$ such that its elements are given thus

$$u_{ij}^{(p)} = \begin{cases} +1 & \text{true dark ink pixels} \\ -1 & \text{false white space pixels} \end{cases} \tag{1}$$

Consider the 2D image as a vector of length N^2 formed by concatenating the columns of the image[4].

$$\mathbf{u} = (u_{11}, u_{21}, ..., u_{N1}, u_{12}, ..., u_{N2}, ..., u_{N^2}) \tag{2}$$

From this subtract the N^2 mean of all images $\bar{\mathbf{u}}$ and insert the result into the columns of the compound image matrix $\mathbf{U}$. The covariance matrix, $\mathbf{R}$, gives the mean, over all images in the ensemble, of all the N^2xN^2 interpixel correlations, and as such, statistically describes how handwritten character images vary. The matrix $\mathbf{R}$ is symmetric and is formed as the outer product of P image vectors.

$$\mathbf{R} = \mathbf{U}\mathbf{U}^T \tag{3}$$

The covariance matrix $\mathbf{R}$ has N^2 eigenvectors as the columns of Ψ defined in the equation

$$\mathbf{R}\Psi = \Psi\Lambda \tag{4}$$

where the only non zero elements of Λ are the eigenvalues λ_i on its diagonal. The eigenvectors are the directions of maximum variance in the N^2 space and form a complete orthonormal set termed

[4] Any consistent ordering is sufficient.

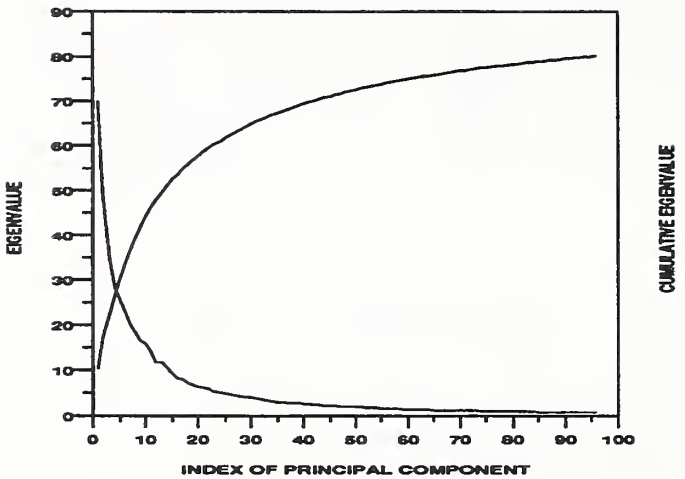

Figure 1: The first ninety six eigenvalues of the covariance matrix shown in figure 3. Note that the cumulative eigenspectrum quickly rises above 70 % of its total.

the principal axes[5] of a hyperellipse in that space. The eigenvalues $diag(\Lambda)$ define the statistical length of these axes as defined by the image data set; thus the first column of Ψ corresponding to the largest eigenvalue is the major axis. Any set of N^2 vectors as the columns of a matrix U can be expressed as a linear combination of the basis vectors:

$$\mathbf{U} = \Psi\mathbf{V} \tag{5}$$

where the inversion of this formula, V, defines the Karhunen Loève Transform, the elements of which are the projection of the image vector onto the principal axes:

$$\mathbf{V} = \Psi^T\mathbf{U} \tag{6}$$

The first sixteen eigenvectors of a covariance matrix are shown in figure 4.

2.2 Data Decorrelation

The KL transform vectors in the columns of V are to be used as input to some classifier. The variance of the KL coefficients themselves is of interest.

$$\mathbf{R_V} = \mathbf{V}\mathbf{V}^T = \Psi^T\mathbf{U}\mathbf{U}^T\Psi = \Lambda \tag{7}$$

That is, the covariance matrix of the KL transforms is diagonal indicating that, by design, the Karhunen Loève Transform perfectly decorrelates the input image data. Specifically the variances of the KL coefficients, σ_i^2, are the respective eigenvalues of the original covariance matrix.

$$\sigma_i = \sqrt{\lambda_i} \tag{8}$$

2.3 Image Reconstruction

The eigenvalue spectrum of figure 1 falls off quickly. The percentage of the variance attributable to L principal components is given on the right hand axis. Geometrically the hyperellipse defined by the eigensolutions of R has extent in only a few directions; along these axes the eigenvalues are large

[5] The Karhunen Loève transform is also known as the method of principal components or the Hotelling transform.

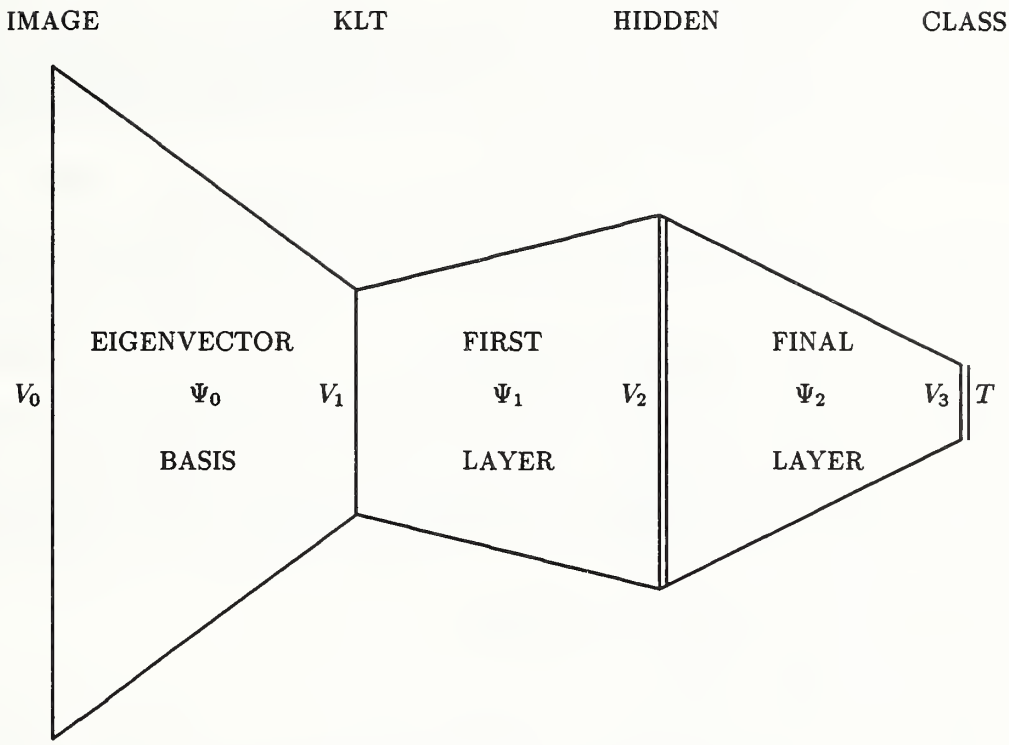

Figure 2: Recognition Architecture. All weight layers are fully connected. The eigenvectors are obtained a priori to the training of the subsequent layers.

indicating that the characters have a large range in their values. A small eigenvalue indicates low variance and is therefore of little utility in describing the differences among the different characters.

Any image is exactly a linear combination of a complete set of conformant orthonormal basis vectors. The KL transform acheives, among the unitary transforms, the maximum energy compaction in a subset of its coeffiecients *on average over those images that define the basis.* For any *given* image the Singular Value Decomposition will achieve maximum energy compaction. If an incomplete basis is used then a reduction in dimensionality, analogous in the Fourier domain to low pass filtering, corresponds to removing spurious variance in the original characters. The KL transform is optimal at image reconstruction in the context of minimal mean square error between original and filtered images. That is, the reconstruction error for the whole image ensemble is merely the sum of the eigenvalues corresponding to the eigenvectors that were not used in the superposition. The cumulative eigenvalue spectrum shows the sum of the eigenvalues as a percentage of the trace of the covariance matrix. Thus is if the dimension of the transform space is 3% of the image space then there is a mean 20% error in the reconstruction of the ensemble. Thus we may inexpensively dispense with the low variance low information coefficients.

2.4 The Layered Perceptron Network

The two weight layer[6] perceptron nonlinearly classifies KL feature vectors. With the evolved KL feature extraction, the network may be regarded as the three layer character classifier of figure 2.

[6]The author has elected to resolve the ambiguity in counting either layers of weights or layers of neurons, pervasive throughout the perceptron literature, by adopting the more minimalist standard of counting weight layers.

The first set of weights Ψ_0 is the *pre-trained* incomplete eigenvector basis set of equation 5. The latter perceptron weight layers, also fully interconnected, are trained using the conjugate gradient algorithm outlined in section 3.6.

All the Karhunen Loève transform vectors are propagated through the network together and the weights are updated. This is *batch mode* training. The use of different subsets of the training patterns to calculate each weight update is known as *on line* training. It is not used in this investigation. Formally the forward propagation is represented as:

$$\mathbf{V_1} = \mathbf{\Psi_0^T V_0} \Longrightarrow \mathbf{V_2} = \mathbf{f}\left(\mathbf{\Psi_1^T V_1}\right) \Longrightarrow \mathbf{V_3} = \mathbf{f}\left(\mathbf{\Psi_2^T V_2}\right) \tag{9}$$

where the network nonlinearity is introduced by squashing all activations with the usual sigmoid function $f(x) = (1 + e^{-x})^{-1}$.

2.5 Classification of Unknown Characters

The linear superposition of a complete set of orthogonal basis functions will describe an arbitrary image. However the whole motivation for using the KLT is to reduce the dimensionality of the feature space by adopting an incomplete basis, i.e. the leading principal components. Only images that resemble the original training characters are adequately representable by the incomplete basis. It is important therefore that the eigenvectors are obtained from a statistically large sample.

3 Experimental Implementation

3.1 Text Page Image Database

The National Institute for Standards and Technology has produced three reference databases on compact disc. The first CD [21] was released in June 1990 and contains the compressed images of 2100 Handwriting Sample Forms, each from one writer. Each form includes twenty fields of handprinted digits or alphabetics. The intended characters are printed above each field such that the completed forms contain unconstrained digits of known class. Of the 270000 characters available on this CD, the first 102340, obtained from 944 different writers, were used for experimentation.

3.2 Page Segmentation

Isolated characters from these forms were obtained after field isolation from a segmentation code that uses an adaptive rule enhanced spatial histogram technique due to Wilkinson [20]. With some inevitable error, isolated 32 pixel square binary images of field centered, size normalized, characters are produced. These characters have been individually verified by a human operator.

3.3 Shear Transformation

To aid recognition, a shear transform is applied to the images. The result is consistently upright, less slanted character images. This approximately obviates the need for the classifier to be rotationally invariant. The shear amount is determined simply by pixel location at the top and bottom of the image yielding a virtual slanted line between them. The rows of the image are shifted horizontally to make the line vertical. This transformation is formally represented as

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & \cot\theta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{10}$$

where the angle θ is the acute interior angle of the line with the horizontal. Figure 6 shows the mean digit images, by class. At left is the raw isolated image; to its right is the mean of all the sheared characters. At bottom center is the mean of *all* characters and its sheared counterpart.

3.4 Covariance Matrices of Binary Images

The efficient calculation of the correlation of binary images is merely the mean of the logical NXOR of the two matrices formed by replicating, as rows and columns, the binary vector. This matrix is the correlation matrix and is converted to the covariance by subtraction of the outer product of the mean vector image. Given 32 pixel square characters, this matrix is held as a 32 bit floating point 1024 by 1024 element array.

3.5 Eigenvectors of the Covariance Matrix

Only the leading eigenvectors are needed. These are obtained by Givens Householder tridiagonalization, application of Power method iteration [19] to find the eigenvalues and eigenvectors, and final rotation of these eigenvectors back to those of the original matrix. This operation is an infrequent expense since the basis, once made, can be used ad infinitum for KL expansion. The eigenvector calculation used 75753 characters drawn almost equally from all ten classes.

3.6 Conjugate Training Algorithm for Feed Forward Neural Networks

Backpropagation [16] is the common method for training multilayer perceptron networks. Essentially it implements a first order minimization of some error objective. The algorithm has the disadvantages that convergence is slow [3] and that there are, in the usual implementation [16], two adjustable parameters, η and α, that have to be manually optimized for the particular problem.

Conjugate gradient methods have been used for many years [5] for minimizing functions, and have recently [8] been discovered by the neural network community. The usual methods require an expensive line search or its equivalent. Møller [13] has introduced a scaled conjugate gradient method; instead of a line search, an estimate of the second derivative along the search direction is used to find the approximate minimum. In both backpropgation and scaled conjugate gradient, by far the most time-consuming part of the calculation is done by the forward error and gradient calculation. In backpropagation this is done once per iteration. Although the scaled conjugate gradient method does this twice per iteration (but occasionally only once), the factor of two overhead is algorithmically negligible since convergence is an order of magnitude faster [13] [1].

3.7 Training and Testing

Except for the third experiment in which the number of training exemplars is varied, the number of training patterns was fixed at 7400. This set was comprised approximately of equal numbers of each class '0' through '9'. Different starting weights yield alternative minima corresponding to a distribution of network performance. Training was performed using tens of uniformly distributed (on the range [-0.5,+0.5]) initial random weights sets. This is insufficient to provide robust statistics but an idea of the variability is obtained. The target activations were 0.0 for all nodes except for a 1.0 on the node representing the given class. The objective function included a regularization [9] term, the square weight vector length.

Testing used the KL transformation of 25585 characters obtained from different writers. This set was disjoint from the training set. The characters from which they were obtained were not used in the calculation of the covariance matrix or its eigenvector basis set. Classification involves a single forward pass through a set of weights. The true classes are known a priori so that the generalization properties of the classifier are obtained.

3.8 Hardware support

The efficient parallel implementation of neural networks in hardware and software is an active area of current research. The motivation is pragmatic; faster training allows larger networks and training

sets to be evaluated. Efficient matrix multiplication, inherent in layered perceptrons, on parallel machines is very architecture specific giving rise to a vast literature on the subject. On an array processor the *outer product* [1] [4] is superior. The AMT DAP[7] possesses a two dimensional array of tightly coupled SIMD processors connected by a high bandwidth bus.

4 Experimental Results

4.1 Dependence on Basis Space Dimension

The graphs of figure 7 show the dependence of training and testing performance for two layer perceptrons on the number of KL coefficients used as feature inputs. The networks that were used in these tables use 32 and 48 hidden nodes respectively. The larger network is clearly superior in learning and classification.

That the recognition does not decline significantly from its maximum at 48 input features as the number of inputs is increased, indicates that the network is capable of ignoring redundant features. This is consistent with Martin and Pittman's work [12]; low variance pixels contain little information and are weighted accordingly. The graphs of figure 7 are averages over 20 different runs using different starting random weight sets. The use of more hidden nodes aids generalization for any sufficient number of KLT inputs and gives better training although, if the number of inputs is large, the number of hidden nodes is increasingly irrelevent. The class is inferred using a *winner takes all* strategy as the index of the highest activation neuron. The activation can be taken as a confidence with which the network asserts its hypothesis and this allows rejection of classifications on the basis of the output activations vector. For example, table 9 and its graph rejects a pattern as unknown if the highest activation is below some threshold. The final result is that the use of more hidden nodes allows the network to train and generalize more successfully.

4.2 Dependence on Number of Training Prototypes

Thirty two KL coefficients of a fixed 15000 patterns were classified by networks trained on up to 40000 patterns. Runs were repeated over at least nine different initial weight sets. Each network had 32 hidden and 10 output units. The results are summarized in the graph of figure 10.

As the number of training exemplars rises the network is less able to learn them. Simultaneously the number that are classified correctly increases. This convergence is also exhibited by the mapping error at the end of training and in testing. With only 32 input features and 32 outputs the network attains 96.1% recognition. From the experiments detailed above it is apparent that more inputs and hidden nodes should be used. Computational limits restricted the number of training exemplars to 40000. The curves of figure 10 indicate that more should be used.

5 Conclusions

The principal components of a training character ensemble form an self-organized basis for feature extraction. The Karhunen Loéve transform is an optimally compact salient linear representation of an image ensemble. It allows character recognition to be performed efficiently and effectively to levels comparable to those of similar studies such as Martin and Pittman[12], and LeCun et al. More importantly the twenty fold reduction in dimensionality is obtained for handwritten digits recognition. Large low dimensional training sets are then available and generalization is shown to be most dependent on this set size. The method is extensible to arbitrary pattern recognition problems including letter OCR.

[7] Certain commercial equipment is identified in order to adequately specify or describe the subject matter of this work. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment identified is necessarily the best available for the purpose.

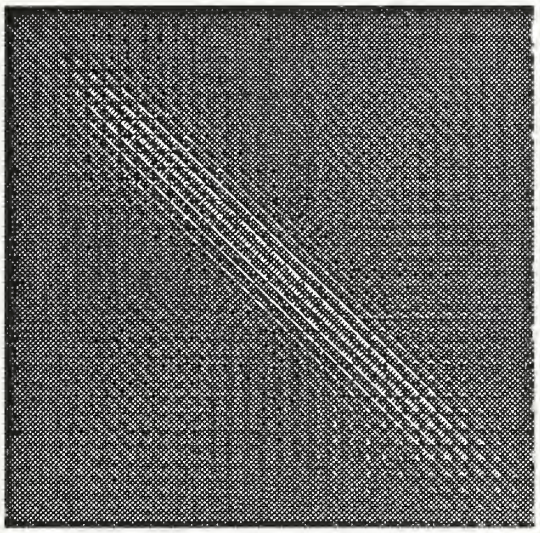

Figure 3: The covariance of 75753 handwritten digits from the sheared 32 by 32 pixel binary images of 940 writers.

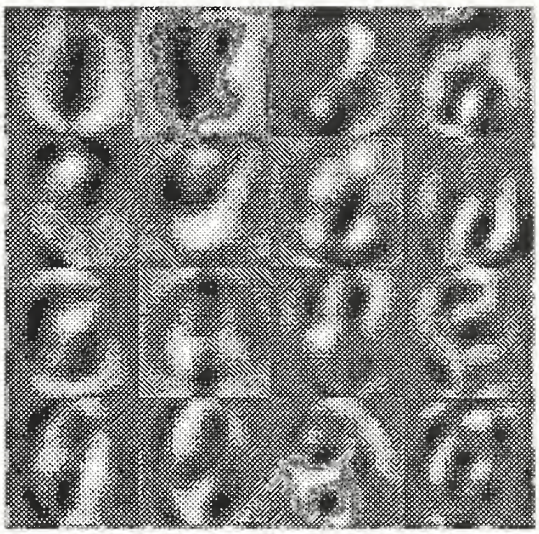

Figure 4: The first sixteen eigencharacters of the covariance matrix of figure 3. They are shown in column major order those with highest eigenvalue first.

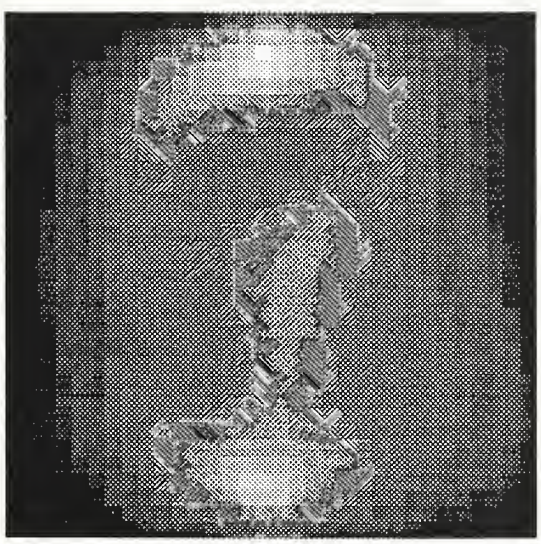

Figure 5: The mean of 75753 handwritten digits from sheared 32 by 32 pixel binary images. The image is zoomed by a zero order hold pixel replication.

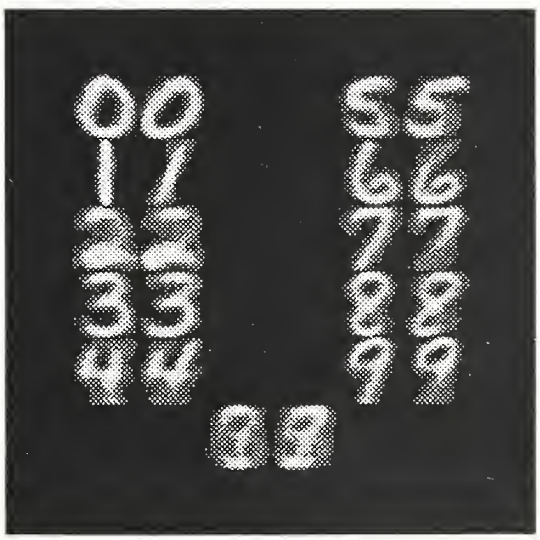

Figure 6: The original and sheared by-class means of 75753 handwritten digits of 32 by 32 pixel binary images from 940 writers. At the bottom is classless mean.

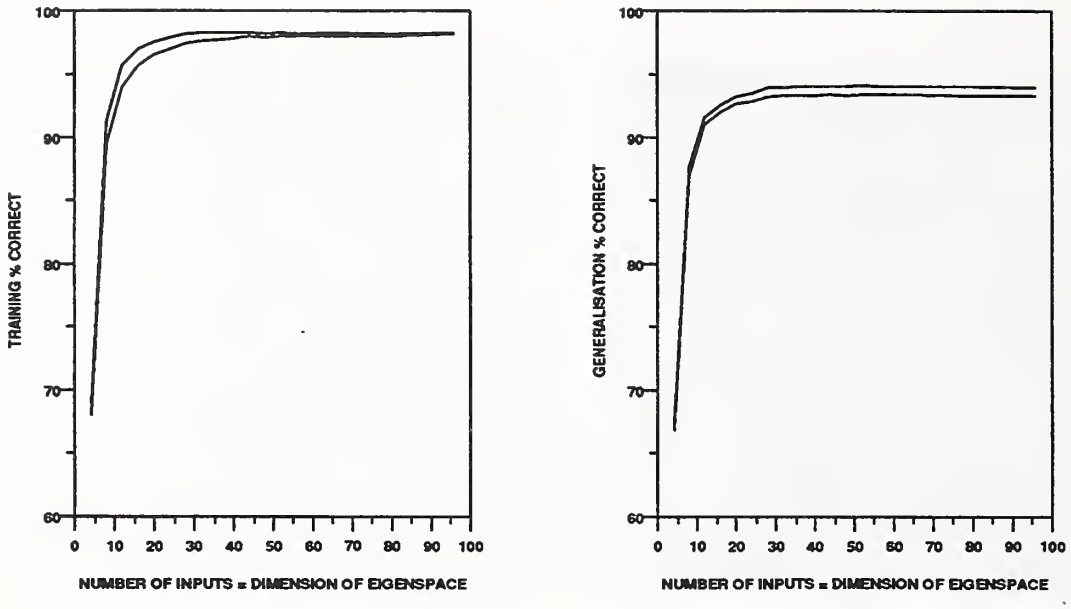

Figure 7: Dependence of recognition accuracy on the number of inputs used. At left the mean percent correct after training on 7400 patterns. At right the results of testing those networks on 25585 new patterns. The higher curves refer to networks with 48 hidden nodes, the lower ones used 32.

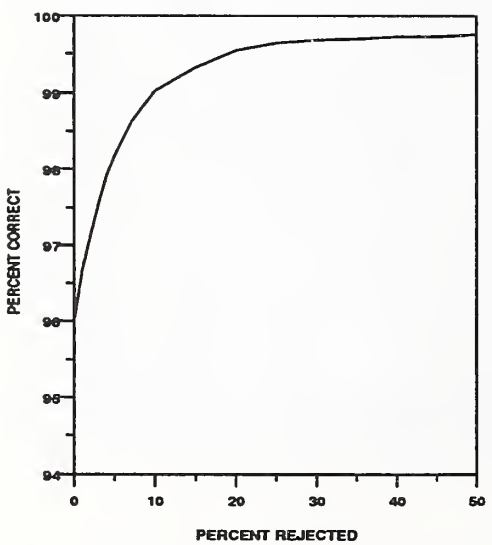

Figure 8: Classification Rejection for 15000 Handwritten Digits

Thresh	*%ok*	*%rej*
0.000	96.05	0.00
0.123	96.67	1.00
0.280	97.18	2.00
0.425	97.58	3.00
0.567	97.93	4.00
0.672	98.18	5.00
0.813	98.62	7.00
0.912	99.03	10.00
0.968	99.33	15.00
0.985	99.55	20.00
0.992	99.64	25.00
0.995	99.69	30.00
0.997	99.70	35.00
0.998	99.73	40.00
0.998	99.73	45.00
0.999	99.75	50.00

Figure 9: Activation Threshold Rejection. 48 inputs 48 hiddens 40000 training and 15000 testing exemplars. At centre is percent classed correctly when the percentage at right of the lowest activation patterns are rejected.

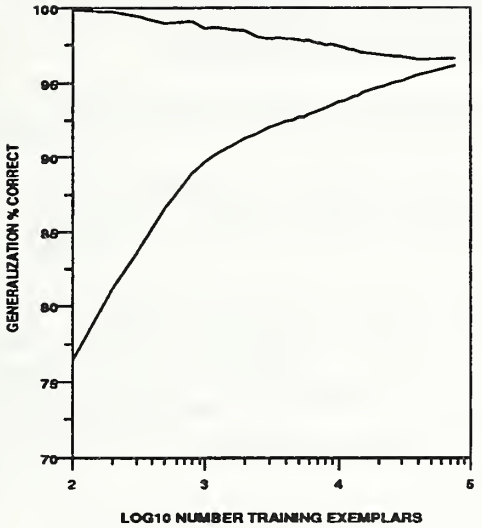

Figure 10: Dependence of training and testing recognition accuracy on the number of training exemplars used. The leading 32 KL coefficients were used.

Training			Testing	
N_{Tr}	P_{ok}	E_{Tr}	P_{ok}	E_{Te}
500	98.97	0.041	86.60	0.149
1000	98.66	0.048	89.69	0.135
2000	98.52	0.052	91.23	0.126
3000	97.97	0.060	92.03	0.122
4000	97.93	0.063	92.43	0.120
6000	97.87	0.066	92.90	0.117
8000	97.54	0.071	93.31	0.114
10000	97.44	0.074	93.77	0.111
12000	97.29	0.076	93.96	0.108
15000	97.05	0.079	94.33	0.105
20000	96.92	0.081	94.69	0.102
24000	96.84	0.081	94.92	0.099
30000	96.79	0.082	95.13	0.097
40000	96.58	0.083	95.56	0.094
76755	96.64	0.083	96.13	0.083

Figure 11: Error and % Correct vs Training Set Size. 32 inputs 32 hidden. Means over 19 initial training weight sets on up to 40000 characters and over one run thereafter. Tested on new 15000.

Classification of 15000 handwritten digits from 312 writers is achieved with 96.5% accuracy using a two layer 48 input, 48 hidden and 10 output unit perceptron architecture trained on 76755 patterns. For a 32 input, 32 hidden and 10 output network trained on 7400 patterns the figure is 93.7%. If the number of input and hidden units is increased to 48 the recognition rate rises to 94.5%.

As the training set size increases a fixed architecture perceptron is increasingly unable to memorize that set but enhances its ability to generalise on unknown patterns. At least 50000 training KL feature vectors are needed for the classifier to classify as well in testing as in training. This applies to both percent classified correctly and to the output objective error.

References

[1] J. L. Blue and P. J. Grother. Training feed forward networks using conjugate gradients. In *Conference on Character Recognition and Digitizer Technologies*, pages 1661–18. SPIE, 2 1992.

[2] John G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20:847–856, 1980.

[3] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[4] P. M. Flanders, R. L. Hellier, H. D. Jenkins, C. J. Pavelin, and S. Van Den Berghe. Efficient high-level programming on the amt dap. *IEEE Proceedings: Special Issue on Massively Parallel Computers*, 79(4):524–536, 4 1991.

[5] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Computer Journal*, 7:149–154, 1964.

[6] M. D. Garris, R. A. Wilkinson, and C. L. WIlson. Analysis of a biolgically motivated neural network for character recognition. In *Proceedings: Analysis of Neural Network Applications.* George Mason University, ACM Press, May 1991.

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

[8] E. M. Johansson, F. U. Dowla, and D. M. Goodman. Backpropagation learning for multi-layer feed-forward neural networks using the conjugate gradient method. *IEEE Transactions on Neural Networks*, 1991.

[9] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems.* Prentice-Hall, Englewood Cliffs, NJ, 1974.

[10] R. Linsker. Self organization in a perceptual network. *Computer*, 21, 1988.

[11] D. J. C. MacKay and K. D. Miller. Analysis of linsker's simulations of hebbian rules. *Neural Computation*, 2:169–182, 1990.

[12] G. Martin and J. Pittman. Recognizing handprinted letters and digits using backpropagation. *Neural Computation*, 3:258–267, 1991.

[13] M. F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 1991.

[14] A. Rojer and E. Schwatz. A multi map model for pattern classification. *Neural Computation*, 1:104–115, 1989.

[15] J. Rubner and P. Tavan. A self organizing network for principal component analysis. *Europhysics Letters*, 10:693–698, 1989.

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, volume I, pages 318–362. MIT Press, Cambridge, MA, 1986.

[17] A. L. Stewart and R. Pinkham. A space variant differential operator for visual sensitivity. *Biological Cybernetics*, 64:373–379, 1991.

[18] 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*, pages I 233–238. IEEE, 7 1991.

[19] S. A. Teukolsky W. H. Press, B. P. Flannery and W. T. Vetterling. *Numerical Recipes*, chapter 11. Cambridge University Press, 1989.

[20] R. A. Wilkinson. Segmenting of text images with massively parallel machines. In *Proceedings October 1991.* SPIE, National Institute of Standards and Technology Gaithersburg Mayland, 1991.

[21] C. L. Wilson and M. D. Garris. Handprinted character database. *National Institute of Standards and Technology*, Special Database 1, 1990.

[22] C. L. Wilson, R. A. Wilkinson, and M. D. Garris. Self organizing neural network character recognition using adaptive filtering and feature extraction. *Advances in Neural Networks*, 3, 1991.

NIST-114A
(REV. 3-90)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION OR REPORT NUMBER
NISTIR 4824

2. PERFORMING ORGANIZATION REPORT NUMBER

3. PUBLICATION DATE
APRIL 1992

4. TITLE AND SUBTITLE

Karhunen Loeve Feature Extraction for Neural Handwritten Character Recognition

5. AUTHOR(S)

Patrick J. Grother

6. PERFORMING ORGANIZATION (IF JOINT OR OTHER THAN NIST, SEE INSTRUCTIONS)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
GAITHERSBURG, MD 20899

7. CONTRACT/GRANT NUMBER

8. TYPE OF REPORT AND PERIOD COVERED

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (STREET, CITY, STATE, ZIP)

10. SUPPLEMENTARY NOTES

11. ABSTRACT (A 200-WORD OR LESS FACTUAL SUMMARY OF MOST SIGNIFICANT INFORMATION. IF DOCUMENT INCLUDES A SIGNIFICANT BIBLIOGRAPHY OR LITERATURE SURVEY, MENTION IT HERE.)

The optimality of the Karhunen Loeve (KL) transform is well known. Since its basis is the eigenvector set of the covariance matrix, a statistical, not functional, representation of the variance in pattern ensembles is generated. By using the KL transform coefficients as a natural feature representation of a character image, the eigenvector set can be regarded as an unsupervised biological feature extractor for a (neural) classifier. The covariance matrix and its eigenvectors are obtained from 76753 handwritten digits. This operation is a unique expense; once the basis set is calculated, it forms a linear first layer of a three weight layer feed forward network. The subsequent nonlinear perceptron layers are trained using a scaled conjugate gradient algorithm that typically affords an order of magnitude reduction in computation over the ubiquitous back-propagation algorithm. In conjunction with a massively parallel computer, training is expedited such that tens of initially different random weight sets are trained and evaluated. Increase in training set size (up to 76755 patterns) gives less accurate learning but improved generalization on the fixed disjoint test set. A neural classifier is realized that recognizes 96.1% of 15000 handwritten digits from 944 different writers. This recognition is attributed to KL transform energy compaction optimality.

12. KEY WORDS (6 TO 12 ENTRIES; ALPHABETICAL ORDER; CAPITALIZE ONLY PROPER NAMES; AND SEPARATE KEY WORDS BY SEMICOLONS)

feature extraction; karhunen loeve transform; neural networks; OCR; supervised learning

13. AVAILABILITY

[X] UNLIMITED

[ ] FOR OFFICIAL DISTRIBUTION. DO NOT RELEASE TO NATIONAL TECHNICAL INFORMATION SERVICE (NTIS).

[ ] ORDER FROM SUPERINTENDENT OF DOCUMENTS, U.S. GOVERNMENT PRINTING OFFICE, WASHINGTON, DC 20402.

[X] ORDER FROM NATIONAL TECHNICAL INFORMATION SERVICE (NTIS), SPRINGFIELD, VA 22161.

14. NUMBER OF PRINTED PAGES
15

15. PRICE
A02

ELECTRONIC FORM