Methods for Enhancing Neural Network Handwritten Character Recognition

M. D. Garris, R. A. Wilkinson, and C. L. Wilson
Advanced Systems Division
National Institute of Standards and Technology,
Gaithersburg, MD 20899

International Joint Conference on Neural Networks, Volume I
IEEE, Seattle, July 1991

Abstract

An efficient method for increasing the generalization capacity of neural character recognition is presented. The network uses a biologically inspired architecture for feature extraction and character classification. The numerical methods used are, however, optimized for use on massively parallel array processors. The method for training set construction, when applied to handwritten digit recognition, yielded a writer-independent recognition rate of 92%. The activation strength produced by network recognition is an effective statistical confidence measure of the accuracy of recognition. A method of using the activation strength for reclassification is described which when applied to handwritten digit recognition reduced substitutional errors to 2.2%.

1.0 Introduction

This paper uses a three part method for writer-independent digit recognition. First, character images are used to calculate least squares optimized Gabor components. For the digit recognition problem, 32 Gabor basis functions are used. Second, these coefficients are used as input feature vectors to a classification network trained using back-propagation learning. Finally, the activation strengths of the network are used as first-order Bayesian statistics for the separation of substitutional errors.

The effectiveness of this method is strongly affected by the nature of the training set used. A new method of training set construction is presented which is based on measuring writer variance. This method is shown to increase the generalization ability of a neural character recognition model.

2.0 Network Architecture

The usual method for designing character recognition systems has been top down. Both special purpose hardware [1] and software [2] approaches have been used on the character recognition problem with promising results. A set of features and a method of feature extraction are selected and the resulting classification problem is solved by a neural network. In this work, we have taken a different approach. The general form of input receptor fields which are used in tasks such as binocular vision by vertebrates has been modeled using parallel Gabor functions [3]. The output of these receptor fields is coupled to small networks for detection of position [4]. In this work, an approximation to this model was constructed as shown in Figure 1 and used for handwritten digit recognition.

The model was not specifically designed for character recognition and could be taught any set of images which could be represented by the Gabor functions. Gabor functions are well suited to this application because they allow reasonable quality image reconstruction with a small number of basis functions.

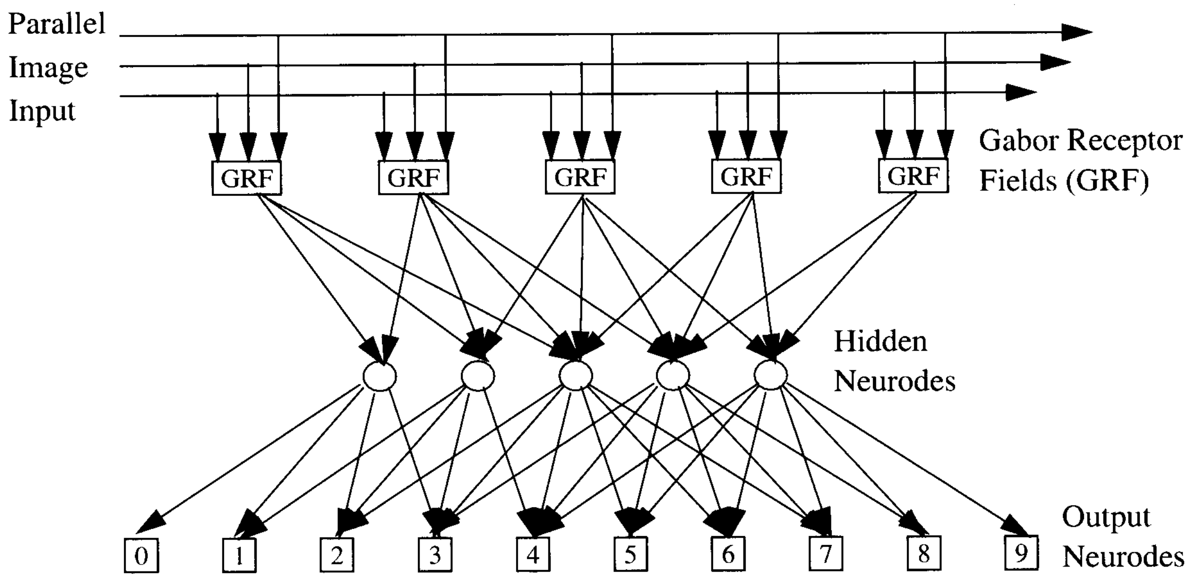

FIGURE 1. Layout of the combined network for feature detection and classification. Most connections are not shown for clarity.

The network in Figure 1 uses parallel processing for the feature extraction process. An image is transmitted to a set of Gabor receptor field (GRF) modules and the combined optimal response is determined by least squares[5]. This method of feature extraction is clearly not biological but is well understood in terms of conventional numerical processing. The outputs of the GRF's are the inputs to a neural network trained using back-propagation learning. The activation of the output neurodes of the back-propagation network are used to classify the input images. Both of the methods used in this paper [6,7] have been used for neural network image processing and recognition applications but have not been applied together previously for handwritten character recognition applications.

Gabor functions are used for input image feature extraction. These functions reduce random image noise and smooth irregularities in image structure by acting as spatially localized low-pass filters. John Daugman [6] has used Gabor functions for image compression and image texture analysis. The most important considerations for character feature extraction applications are that the Gabor functions provide the minimum combination of uncertainty in position and spatial frequency resolution[6] and that the profiles of Gabor functions match the visual receptor field profiles of mammalian eyes[6]. For handwritten character recognition applications, only 32 8-bit Gabor filter coefficients are required to approximate a 1024 pixel image. Image reconstruction using Walsh functions of equivalent quality requires over 100 basis functions [2]. Decreasing the number of required basis functions reduces the number of connections needed in the classification network.

3.0 Training Set Construction

Back-propagation networks provide a very effective method for performing supervised nonlinear classifications [7]. This learning can be enhanced by presenting the network examples which exploit the boundaries of n-dimensional feature space. Presenting a network with examples of this nature provides increased opportunity for learning the functional relationships governing these boundaries. This strategy for choosing training prototypes is supported by principal component analysis in statistics where the eigenvalues of a covariance matrix are used to locate the direction of maximum variance in feature space[8]. However, deriving principal components by either conventional numerical methods[9] or neural network methods[10] is computationally expensive. A new method of feature ranking has been developed assuming the vector components are statistically independent. Studies have been conducted where the covariance matrix has been formed and diagonalized resulting in 16 significant eigenvalues. This demonstrates that the 32 Gabor function coefficients are not statistically independent. However, the ranking process on the reduced basis set would be substantially the same as the method presented below.

Given a set of feature vectors, f_{ij}, where i is the vector index and j is the component index, a mean vector, m_j, can be calculated as:

$$m_j = \frac{1}{p}\sum_{i=0}^{p} f_{ij} \qquad 0 \le j < n \tag{1}$$

A square of the distance from each feature vector to the mean vector is defined as:

$$d_i = \sum_{j=0}^{n} (f_{ij} - m_j)^2 \qquad 0 \le i < p \tag{2}$$

The list, d_i, can then be sorted, forming a list of feature vector indices ranked according to their distance from the mean vector. This technique was used to develop a training set of Gabor feature vectors derived from handwritten characters.

A subset of the NIST Handprinted Character Database was segmented producing a database of over 3,000 handwritten digits collected from 49 different writers[11]. Figure 2 shows a small subset of the characters found in this database. These characters were machine segmented and then visually verified by a human. These character images are a reasonable representation of the types of variations found across the 2,100 writer population of the entire NIST Handprinted Character Database and includes segmentation errors.

FIGURE 2. Sample of segmented handwritten characters.

The input vectors to the network are each composed of 32 Gabor coefficients produced by presenting each character image to a parallel image reconstruction algorithm. A set of Gabor feature vectors was selected for network training from the entire character database using equations (1) and (2). Ten lists of Gabor feature vectors were compiled, one for each character class, 0 through

9, so that each list contained all the examples of a specific character class from all 49 writers. The character class lists were sorted in descending order of variance. Character-writer lists were created from the ranked character lists in the order of each writer's most extreme vector. The position of each writer within each character-writer list was then recorded with a 1 assigned to the first writer found and then incrementing to a 49 for the last writer found. Therefore, each writer was assigned 10 different position values corresponding to their positions in the 10 character-writer lists. The 10 position values for each writer were summed and then used to approximate the overall variance contained within each writer's characters.

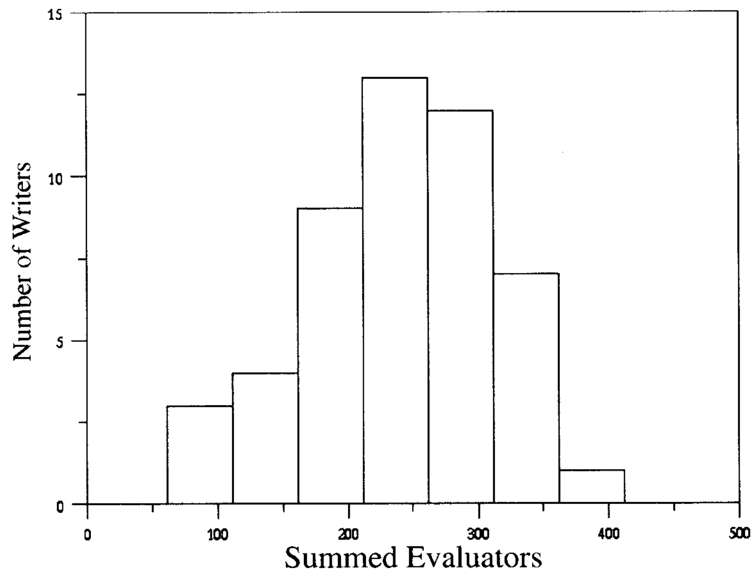

FIGURE 3. Histogram of summed evaluators for 49 writers.

The distribution of the 49 writers' summed evaluators is plotted as a histogram in Figure 3. The writers plotted farthest left have relatively high variance across their characters, whereas those writers plotted farthest right have relatively no variance. From this set of 49 writers, 27 were selected for training creating an evaluator distribution biased towards writers with high variance, but with an adequate number of average writers to help stabilize the network. The remaining 22 writers were used for testing. The standard deviation of the evaluators from all the writers is 71.26, whereas the standard deviation from just those writers in the training set is 82.02. The standard deviation of the evaluators from the writers in the test set is 54.82. This selection process enabled a training set of 2,000 prototypes and a test set of 1,434 test examples to be built.

4.0 Results

Epoch-based back-propagation training was conducted using the 27-writer training set. A three-layered feedforward network was used having 32 input neurodes, 15 hidden neurodes, and 10 output neurodes. A learning rate of 0.001 and a momentum factor of 0.5 were used, and the network was permitted to train for 2000 epochs. Presenting the training set to the final weights of the network showed 91.6% of the training prototypes had been learned correctly. When presented with the test set of 1,434 examples from 22 writers not used in training, the network achieved 92.1% recognition. The highest of the 10 output neurode activations determined the network classification. This demonstrates the network's ability to generalize for truly writer-independent recognition.

In addition to raw recognition results above, a mean positive activation value and a mean negative activation value were calculated from the network responses. These statistics are derived by first scoring all winning output neurode activations against their target values and then separating them into two distinct categories or distributions, those network responses which are correct and those which are incorrect.

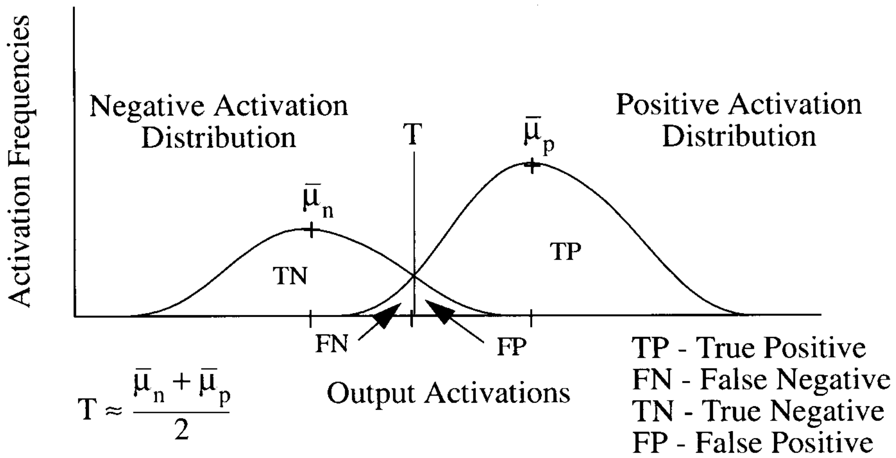

FIGURE 4. Threshold Classifier based on two Gaussian distributions.

Two general observations have been made from studying these mean activations[12]. As the mean positive activation increases, the network performance increases. Also, as the distance between the mean positive and mean negative activations increases, the network performance increases. If the positive and negative activation distributions were Gaussian in shape, then Figure 4 could be used to visualize these generalizations. In Figure 4, the negative activation distribution is shown on the left with a mean value of $\bar{\mu}_n$, and the positive activation distribution is to the right with a mean value of $\bar{\mu}_p$. A threshold, T, is included in Figure 4 which optimally separates one distribution from the other and can be approximated by calculating the midpoint between $\bar{\mu}_n$ and $\bar{\mu}_p$. Using this threshold in place of the *a priori* knowledge, which was used to generated the two distributions, results in reclassifying all activations to the right of the threshold as correct classifications and all activations to the left as rejected classifications. The system responses when evaluated fall into one of four categories, true positive, false negative, true negative, and false positive.

The network reported in this work when presented the 22-writer test set produced a mean positive activation of 0.8408 and a mean negative activation of 0.4628. By applying a threshold at the midpoint of 0.6518, a true positive recognition rate of 80.5% was achieved with only a 2.2% false positive, substitutional, error rate. The remaining 17.2% were classified as unknown and therefore rejected. This demonstrates that the activation strength of the network can be used as an effective first-order confidence level.

5.0 Conclusions

An improved method of training set construction for handwritten character recognition has been presented. The GRF features used for character recognition are shown to be effective for developing training sets which have improved generalization capability. The method used for training set construction is similar to, but much less computationally costly, than the biologically motivated method of principal component analysis [13]. Using this method, handwritten writer-independent digit recognition with 92% accuracy has been demonstrated. A statistical method for confidence measurement allows substitutional errors to be reduced to 2.2% while retaining 80.5% recognition accuracy. Parallel GRF calculations take 13.7 ms per image and network classification on a serial computer takes 19 ms.

Bibliography

[1] L. D. Jackel, H. P. Graf, W. Hubbard, J. S. Denker, D. Henderson, and Isabelle Guyon, "An Application of Neural Net Chips: Handwritten Digit Recognition," IEEE International Conference on Neural Networks, San Diego, Vol. II, pp. 107-115, 1988.

[2] A. Rajavelu, M. T. Musavi, and M. V. Shirvaikar, "A Neural Network Approach to Character Recognition," *Neural Networks*, **2**, pp. 387-393, 1989.

[3] C. L. Wilson, "A New Self-Organizing Neural Network Architecture for Parallel Multi-Map Pattern Recognition - FAUST," in preparation.

[4] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman, "Stereoscopic Depth Discrimination in the Visual Cortex: Neurons Ideally Suited as Disparity Detectors," *Science*, **249**, pp. 1037-1041, 1990.

[5] C. L. WIlson, R. A. Wilkinson, and M. D. Garris, "Self-Organizing Neural Network Character Recognition on a Massively Parallel Computer," *Proc. of the IJCNN*, **II**, pp. 325-329, June 1990.

[6] J. G. Daugman, "Complete Discrete 2-D Gabor Transform by Neural Networks for Image Analysis and Compression," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **36**, pp. 1169-1179, 1988.

[7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Parallel Distributed Processing, Volume 1: Foundations*, edited by D. E. Rumelhart, J. L. McClelland, et al, MIT Press, Cambridge, pp. 318-362, 1986.

[8] M. Kendall, *Multivariate Analysis*, Hafner, New York, pp. 13-29, 1975.

[9] J. Ortega, "The Givens-Householder method for symmetric matrices," *Mathematical Methods for Digital Computers*, **II**, edited by A. Ralston and H. S. Wilf, Wiley, New York, pp. 94-115, 1967.

[10] J. Rubner and K. Schulten, "Development of Feature Detectors by Self-Organization," *Biological Cybernetics*, **62**, pp. 193-199, 1990.

[11] C. L. Wilson and M. D. Garris, "Handprinted Character Database," *NIST Special Database 1*, **HWDB**, April 18, 1990.

[12] M. D. Garris, R. A. Wilkinson, and C. L. Wilson, "Analysis of a Biologically Motivated Neural Network for Character Recognition," *Analysis of Neural Net Applications Conference*, May 30, 1991.

[13] R. Linsker, "Self-Organization in a Perceptual Network," *Computer*, **21**, pp. 105-117, 1988.