High accuracy handwritten Chinese character recognition using quadratic classifiers with discriminative feature extraction

Cheng-Lin Liu

High Accuracy Handwritten Chinese Character Recognition Using Quadratic Classifiers with Discriminative Feature Extraction

Cheng-Lin Liu
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences
PO Box 2728, Beijing 100080, P.R. China
E-mail: liucl@nlpr.ia.ac.cn

Abstract

We aim to improve the accuracy of handwritten Chinese character recognition using two advanced techniques: discriminative feature extraction (DFE) and discriminative learning quadratic discriminant function (DLQDF). Both methods are based on the minimum classification error (MCE) training method of Juang et al. [7], and we propose to accelerate the training process on large category set using hierarchical classification. Our experimental results on two large databases show that while the DFE improves the accuracy significantly, the DLQDF improves only slightly. Compared to the modified quadratic discriminant function (MQDF) with Fisher discriminant analysis, the error rates on two test sets were reduced by factors of 29.9% and 20.7%, respectively.

1. Introduction

Handwritten Chinese character recognition (HCCR) is difficult due to the large category set, wide variability of writing styles, and the confusion between similar characters. It has been attacked intensively from 1980s, and many effective methods have been proposed. Some important techniques, including directional feature extraction, nonlinear normalization, and modifications of quadratic classifiers, have contributed to today's high accuracies on handprinted characters. Even higher accuracies are desired for both handprinted and unconstrained handwritten characters, however.

As a variant of quadratic classifiers, the modified quadratic discriminant function (MQDF) of Kimura et al. [1] has yielded superior performance in large character set recognition [2]. The MQDF is often combined with Fisher discriminant analysis (FDA), which reduces the dimensionality of features with little loss of accuracy. Other improvements were given by compound quadratic discriminant functions for discriminating character pairs [3], and asymmetric Gaussian discriminant functions [4]. For these classifiers, high accuracies are attributed to the nearly Gaussian distribution of each class. Discriminative classifiers, like neural networks and support vector machines, encounter difficulties in training with large category set and large number of samples.

We aim to improve the accuracy of HCCR using two advanced techniques: discriminative feature extraction (DFE) [5] and discriminative learning quadratic discriminant function (DLQDF) [6]. With DFE, the subspace axes are learned simultaneously with the parameters of the underlying classifier by iterative optimization of the minimum classification error (MCE) criterion of Juang et al. [7]. The subspace learned hereby is expected to have better separability for similar characters than that learned by parametric discriminant analysis. The DLQDF is a discriminatively updated version of MQDF, and was shown to improve significantly the accuracy of handwritten numeral recognition [6]. For training DLQDF on large category set, we alleviate the heavy computation using hierarchical classification.

The DFE strategy has shown success in large character set recognition with simple classifier structures like the nearest prototype classifier (see [8] and references therein). Since the training of DFE with quadratic classifiers is complicated, we learn a subspace using DFE with the nearest prototype classifier, then use the MQDF or DLQDF on the learned subspace for classification. In a previous work [9], the DLQDF was used with FDA only, and the eigenvectors of each class were not updated discriminatively.

We have evaluated the effects of DFE and DLQDF on two large databases, ETL9B database and CASIA (Institute of Automation, Chinese Academy of Sciences) database. Our results show that when compared with the MQDF combined with FDA, the DFE improves the accuracy significantly while the DLQDF improves it only slightly. Overall, the error rates were reduced by factors of over 20%.

2. Chinese Character Recognition System

The block diagram of our HCCR system is shown in Fig. 1. In pre-processing, the character image is normalized to a standard size. After feature extraction, the feature vector $\mathbf{x} = [x_1, \ldots, x_d]^T$ is projected onto a low dimensional subspace: $\mathbf{z} = \Phi^T \mathbf{x} = [z_1, \ldots, z_m]^T$ ($m < d$, Φ is the transformation matrix composed of the sub-

space axes). The reduced feature vector $\mathbf{z}$ is then fed into the classification module to output class codes and corresponding distance/similarity scores. The classification module consists of a coarse classifier for candidate selection and a quadratic classifier, MQDF or DLQDF computed on the candidate classes only. The coarse classifier also has two hierarchies: cluster-based class-group classification and prototype-based classification within selected groups [8]. The nearest prototype classifier alone is not fast enough.

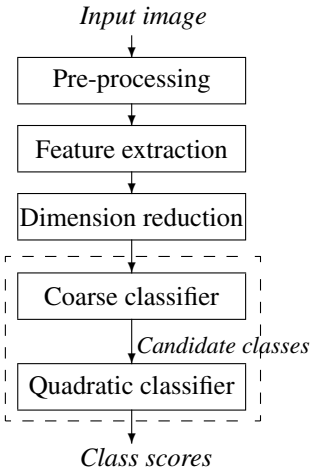

Figure 1. Diagram of the HCCR system

For character image normalization, we use a global curve fitting-based normalization method, named bi-moment normalization [10], which yields comparable performance with line density-based nonlinear normalization at simpler computation. We transform the binary input image into a gray-scale normalized image of 64×64 pixels by continuous pixel mapping and extract gradient direction feature [11]. A 512-D feature vector is obtained by sampling 64 values from each of eight direction planes with Gaussian blurring.

3. Discriminative Feature Extraction

The basis of subspace Φ is learned on a sample dataset with the aim of optimizing an objective criterion of class separability. The Fisher discriminant analysis (FDA) aims to maximize the between-class scatter while maintaining the compactness of within-class scatter by optimizing the Fisher criterion [12]:

$$\max_{\Phi} tr[(\Phi^T S_w \Phi)^{-1}(\Phi^T S_b \Phi)], \tag{1}$$

where S_w and S_b are the within-class scatter matrix and between-class scatter matrix, respectively. The resulting m basis vectors (columns of Φ) are the eigenvectors of $S_w^{-1} S_b$ corresponding to the m largest eigenvalues. FDA is optimal when the feature densities of all classes are multivariate Gaussian sharing a common covariance matrix, but this generally does not hold.

Discriminative feature extraction (DFE) [5] optimizes the subspace axes to minimize the classification error on the training sample set. The training error is given by a classifier on the feature subspace, and the classifier parameters and subspace axes are optimized simultaneously. For quadratic classifiers like the MQDF, however, the simultaneous optimization of classifier parameters with subspace axes is complicated. Hence, we learn the subspace combined with a nearest prototype classifier, which is computationally feasible on large category set.

Given a training sample set $X = \{(\mathbf{x}^n, c^n)|n = 1, \ldots, N\}$ (c^n is the class label of sample $\mathbf{x}^n$), initial subspace basis $\Phi(0) = [\phi_1(0), \ldots, \phi_m(0)]$ and the prototypes $\Theta(0) = \{\mathbf{m}_i(0)|i = 1, \ldots, M\}$ of M classes, the learning task is to adjust Φ and Θ by minimizing the classification error on X. On the initialized subspace learned by FDA, we set the initial prototypes to the means of projected vectors of samples of each class.

Denoting by $g_i(\mathbf{x})$ as the discriminant function for class ω_i, the misclassification measure can be computed by

$$h_c(\mathbf{x}) = -g_c(\mathbf{x}) + g_{r(c)}(\mathbf{x}), \tag{2}$$

where c denotes the genuine class of $\mathbf{x}$ and $r(c)$ is the closest rival class: $g_{r(c)}(\mathbf{x}) = \max_{i \neq c} g_i(\mathbf{x})$. For nearest prototype classifier, the measure is specified as

$$h_c(\mathbf{x}) = \|\Phi^T \mathbf{x} - \mathbf{m}_c\|^2 - \|\Phi^T \mathbf{x} - \mathbf{m}_{r(c)}\|^2, \tag{3}$$

and is transformed to loss by

$$l_c(\mathbf{x}) = l_c(h_c) = \frac{1}{1 + e^{-\xi h_c}}. \tag{4}$$

The empirical loss on the sample set is

$$L_0 = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{M} l_i(\mathbf{x}^n) I(\mathbf{x}^n \in \omega_i). \tag{5}$$

In our implementation, we added a regularization term (which constrains the deviation of prototypes from class means) to the empirical loss and updated the parameters iteratively by stochastic gradient descent [8].

A major source of computation in MCE training is the search for closest rival class, which was accelerated by cluster-based coarse classification. The cluster prototypes were updated together with the class prototypes and subspace axes. This was shown to improve the tradeoff between the classification accuracy and the number of candidates [8].

4. MQDF and DLQDF

The DLQDF is a discriminatively updated version of the MQDF of Kimura et al. [1]. On a feature vector $\mathbf{x}$, the MQDF (also called quadratic distance below) for a class ω_i is computed by

$$\begin{aligned} d_Q(\mathbf{x}, \omega_i) = \; & \textstyle\sum_{j=1}^{k} \frac{1}{\lambda_{ij}} [\phi_{ij}^T (\mathbf{x} - \mu_i)]^2 + \frac{1}{\delta_i} r_i(\mathbf{x}) \\ & + \textstyle\sum_{j=1}^{k} \log \lambda_{ij} + (d - k) \log \delta_i, \end{aligned} \tag{6}$$

where μ_i is the mean vector of class ω_i , λ_{ij} $(j = 1, \ldots, k)$ are the largest eigenvalues of the covariance matrix and ϕ_{ij} are the corresponding eigenvectors, k denotes the number of principal axes and $r_i(\mathbf{x})$ is the residual of subspace projection: $r_i(\mathbf{x}) = \|\mathbf{x} - \mu_i\|^2 - \sum_{j=1}^{k}[(\mathbf{x} - \mu_i)^T \phi_{ij}]^2$. δ_i is a class-specific or class-independent constant to substitute the minor eigenvalues. We set it to a class-independent constant and optimize its value by holdout cross validation on the training data set.

The MQDF has reduced complexity and can give improved classification compared to the original quadratic discriminant function (QDF). To overcome the non-Gauassianity of probability densities, the parameters of MQDF can be optimized on training samples by optimizing the MCE criterion. The optimized discriminant function is called discriminative learning QDF (DLQDF) [6].

The parameters of DLQDF (mean vectors, eigenvalues and eigenvectors) are iteratively updated on a training sample set to minimize the empirical loss. Taking discriminant function $g_i(\mathbf{x}) = -d_Q(\mathbf{x}, \omega_i)$, the misclassification measure (2) is specified as

$$h_c(\mathbf{x}) = d_Q(\mathbf{x}, \omega_c) - d_Q(\mathbf{x}, \omega_{r(c)}), \tag{7}$$

and is transformed to loss by (4). The empirical loss is summed up over a training sample set as in (5). To constrain the motion of parameters, we added a regularization term related to maximum likelihood (ML) to the empirical loss:

$$L_1 = \frac{1}{N}\sum_{n=1}^{N}[l_c(\mathbf{x}^n) + \alpha d_Q(\mathbf{x}^n, \omega_c)], \tag{8}$$

where $d_Q(\mathbf{x}^n, \omega_c)$ is the quadratic distance between the input pattern and the genuine class and α is the regularization coefficient.

In updating the parameters of DLQDF by stochastic gradient descent, three different learning steps are set for the eigenvalues, mean vectors, and eigenvectors, respectively. We keep the eigenvalues positive by transforming them into exponential functions and keep the eigenvectors of each class ortho-normal by Gram-Schmidt orthonormalization. More details can be found in [6].

For training DLQDF on large category set, the search for the rival class of minimum quadratic distance on each training sample is very expensive. We speed up the training process by computing quadratic distance on candidate classes selected by a two-level hierarchical prototype classifier as for DFE. The class prototype is the adjustable mean of each class, the group prototypes are initialized to be the cluster centers of class means and are updated discriminatively together with the parameters of DLQDF.

We use MQDF or DLQDF on projected features in linear subspace ($\mathbf{z} = \Phi^T\mathbf{x}$ to replace $\mathbf{x}$ in (6)), but do not attempt to optimize the subspace axes together with the parameters of DLQDF because the computation is too complicated. Instead, we learn the discriminative subspace combined with a nearest prototype classifier and use MQDF or DLQDF on this learned subspace.

5. Experimental Results

We experimented on two large databases of handwritten characters, namely, the ETL9B database collected by the Electro-Technical Laboratory of Japan and a database collected by the Institute of Automation, Chinese Academy of Sciences (CASIA). The ETL9B database has been tested in many previous works (e.g., [2, 4, 9]). It contains handwritten samples of 3,036 characters, including 2,965 Kanji characters and 71 hiragana, 200 samples per class. We used 40 samples (first 20 and last 20) from each class for testing and the remaining 160 samples of each class for training. The CASIA database contains handwritten samples of 3,755 Chinese characters of the level-1 set of the standard GB2312-80, 300 samples per class. We used 250 samples from each class for training, and the remaining 50 samples of each class for testing.

We used the MQDF or DLQDF for classification on candidate classes selected by a two-level prototype classifier. We used 220 cluster prototypes for the ETL9B database and 250 cluster prototypes for the CASIA database. For both databases, we tested MQDF with 50 eigenvectors per class on linear subspace of variable dimensionality learned by FDA and DFE, and on a 160-D subspace, we tested MQDF and DLQDF with variable number of eigenvectors per class. Further, the DLQDF was given two variations: the DLQDF1 that does not update eigenvectors and the DLQDF2 that updates all parameters.

Table 1. Error rates (%) on ETL9B using MQDF combined with FDA and DFE

subspace	120	140	160	180	200
FDA	0.91	0.90	**0.87**	0.88	0.90
DFE	0.75	0.70	0.67	0.66	**0.65**

Table 2. Error rates (%) on ETL9B using MQDF and DLQDF on 160-D subspace

	#eigenvector	20	30	40	50
FDA	MQDF	0.92	0.88	**0.87**	0.87
	DLQDF1	0.87	0.81	**0.79**	0.79
	DLQDF2	0.84	0.81	0.81	**0.80**
DFE	MQDF	0.73	0.66	**0.65**	0.67
	DLQDF1	0.67	**0.61**	0.61	0.63
	DLQDF2	0.68	**0.63**	0.63	0.64

The test error rates on the ETL9B database using MQDF on subspace of variable dimensionality are shown in Table 1. The error rate of MQDF without dimensionality reduction is 0.92%, which is inferior than with dimensionality

Table 3. Error rates (%) on CASIA using MQDF combined with FDA and DFE

subspace	120	140	160	180	200
FDA	2.12	2.02	1.99	**1.97**	1.97
DFE	1.82	1.75	1.72	1.70	**1.69**

Table 4. Accuracies on CASIA (%) using MQDF and DLQDF on 160-D subspace

	#eigenvector	30	40	50	60
FDA	MQDF	2.04	**1.98**	1.99	2.01
DFE	MQDF	1.73	**1.68**	1.72	1.80
	DLQDF1	1.65	**1.59**	1.63	1.67
	DLQDF2	1.61	**1.57**	1.58	1.60

reduction. We can see that compared to dimensionality reduction by FDA, the DFE improves the accuracy significantly. On 160-D subspace, the error rate was reduced from 0.87% to 0.67%. The error rates of MQDF and DLQDF on 160-D subspace are shown in Table 2. We can see that on subspace learned by either FDA or DFE, the DLQDF gives lower error rates than the MQDF. The improvement is not so significant as that made by DFE, however. The fact that the error rates of DLQDF1 and DLQDF2 are comparable indicates that discriminative updating of class eigenvectors does not help for handwritten Chinese characters, in contrast to handwritten numeral recognition [6].

The test error rates on the CASIA database are shown in Table 3 and Table 4. For this database, we did not experiment with DLQDF on the subspace learned by FDA. The results of Table 3 agree that DFE improves the accuracy of MQDF significantly compared to FDA. Table 4 shows that the accuracy of MQDF was improved slightly by discriminative learning (DLQDF) and the accuracies of DLQDF1 and DLQDF2 are again comparable.

On the ETL9B database, the error rate of MQDF-FDA, 0.87%, was reduced to 0.61% of DLQDF-DFE. On the CASIA database, the error rate of MQDF-FDA, 1.98%, was reduced to 1.57%. Accordingly, the error reduction rates of two databases are 29.9% and 20.7%, respectively.

6. Conclusion

We proposed to improve the accuracy of handwritten Chinese character recognition using DFE and DLQDF. For both methods, we accelerated the training on large category set using hierarchical classification. Though we circumvented the complication of optimizing the subspace axes and DLQDF parameters simultaneously, the DFE with a nearest prototype classifier improved the accuracy of MQDF and DLQDF significantly. On the subspace learned by either FDA or DFE, the DLQDF improves the accuracy only slightly. Overall, we could achieve error reduction rates of over 20% based on a state-of-the-art system. With acceleration by hierarchical classification, the DFE and DLQDF methods are feasible for even larger category set with over 10,000 classes.

Acknowledgements

This work is supported in part by the Hundred Talents Program of Chinese Academy of Sciences and the Natural Science Foundation of China (grant no.60543004). The author appreciates the proofreading of Prof. George Nagy.

References

[1] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and the application to Chinese character recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, 9(1): 149-153, 1987.

[2] F. Kimura, T. Wakabayashi, S. Tsuruoka, Y. Miyake, Improvement of handwritten Japanese character recognition using weighted direction code histogram, *Pattern Recognition*, 30(8): 1329-1337, 1997.

[3] M. Suzuki, S. Omachi, N. Kato, H. Aso, Y. Nemoto, A discrimination method of similar characters using compound Mahalanobis function, *Trans. IEICE Japan*, J80-D-II(10): pp.2752-2760, 1997.

[4] N. Kato, M. Suzuki, S. Omachi, H. Aso, Y. Nemoto, A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance, *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(3): 258-262, 1999.

[5] A. Biem, S. Katagiri, B.-H. Juang, Pattern recognition using discriminative feature extraction, *IEEE Trans. Signal Processing*, 45(2): 500-504, 1997.

[6] C.-L. Liu, H. Sako, H. Fujisawa, Discriminative learning quadratic discriminant function for handwriting recognition, *IEEE Trans. Neural Networks*, 15(2): 430-444, 2004.

[7] B.-H. Juang, S. Katagiri, Discriminative learning for minimum error classification, *IEEE Trans. Signal Processing*, 40(12): 3043-3054, 1992.

[8] C.-L. Liu, R. Mine, M. Koga, Building compact classifier for large character set recognition using discriminative feature extraction, *Proc. 8th ICDAR*, Seoul, Korea, 2005, pp.846-850.

[9] H. Liu, X. Ding, Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes, *Proc. 8th ICDAR*, Seoul, Korea, 2005, pp.19-23.

[10] C.-L. Liu, H. Sako, H. Fujisawa, Handwritten Chinese character recognition: alternatives to nonlinear normalization, *Proc. 7th ICDAR*, Edinburgh, Scotland, 2003, pp.524-528.

[11] C.-L. Liu, K. Nakashima, H. Sako, H. Fujisawa, Handwritten digit recognition: investigation of normalization and feature extraction techniques, *Pattern Recognition*, 37(2): 265-279, 2004.

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition, Academic Press, 1990.