Error Reduction by Confusing Characters Discrimination for Online Handwritten Japanese Character Recognition

Xiang-Dong Zhou [1], Da-Han Wang [2], Masaki Nakagawa [1], Cheng-Lin Liu [2]
[1] *Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan*
[2] *National Laboratory of Pattern Recognition (NLPR),*
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P.R. China
[1] *{xdzhou, nakagawa}@cc.tuat.ac.jp,* [2] *{dhwang, liucl}@nlpr.ia.ac.cn*

Abstract

To reduce the classification errors of online handwritten Japanese character recognition, we propose a method for confusing characters discrimination with little additional costs. After building confusing sets by cross validation using a baseline quadratic classifier, a logistic regression (LR) classifier is trained to discriminate the characters in each set. The LR classifier uses subspace features selected from existing vectors of the baseline classifier, thus has no extra parameters except the weights, which consumes a small storage space compared to the baseline classifier. In experiments on the TUAT HANDS databases with the modified quadratic discriminant function (MQDF) as baseline classifier, the proposed method has largely reduced the confusion caused by non-Kanji characters.

1. Introduction

Character recognition of a large vocabulary, like Chinese and Japanese, commonly adopts a hierarchical classification scheme for improving both the classification speed and accuracy [1, 2, 3]. By the widely adopted two-stage scheme, the first stage uses a fast coarse classifier for selecting a small subset of candidate classes with the hope of containing the genuine class of input pattern, and the second stage uses a high accuracy classifier for identifying the class of input pattern from the candidate classes [1, 2]. The first-stage classifier usually adopts the Euclidean distance or linear discriminant function, which are computationally simple.

The second stage of large vocabulary character recognition has widely adopted quadratic classifiers, especially, the modified quadratic discriminant function (MQDF) classifier [4]. Quadratic classifiers have the merit that they provide fairly high classification accuracy and the parameters of each class are estimated independently of the other classes. This generative training strategy is fast, but does not provide as high accuracy as discriminative classifiers because the separability between classes is not considered in training. On the other hand, the discriminative training of quadratic classifiers is too complicated for a problem of thousands of classes.

To improve the accuracy of quadratic classifiers, the compound Mahalanobis function (CMF) was proposed [5]. For discriminating a pair of classes (the top two ranks output by the baseline classifier), the CMF projects the class mean difference onto the respective complement subspace of two classes, and the distance on this projected feature is combined with the baseline quadratic discriminant function for discrimination. Compared to the MQDF, the CMF has no extra parameters but the online computation of pair discriminant functions is expensive. The recently proposed LDA (linear discriminant analysis)-based compound distance [6] and the critical region analysis-based pair discrimination [7] further improves the accuracy of CMF at the cost of extra parameters storage. For achieving high enough accuracy, the number of pair discriminators is very large, e.g., over 70,000 pairs for a 3,755-class character set in [6].

A more straightforward strategy for discriminating confusing characters is just to re-classify the input pattern in a subset of confusing classes with a discriminative classifier, such as the neural network in [8]. This method re-classifies the input pattern whenever the top rank class of baseline classifier falls in a confusion set. For high accuracy, the number of confusion set discriminators is not small and each discriminator is complex. So, the additional cost of storage and computation is still considerable.

In this paper, we propose a new method to discriminate confusing characters. First, we form confusing class sets by cross validation using a baseline MQDF classifier. Each confusing set is classified by a logistic regression (LR) classifier [9] with subspace features selected from the existing vectors of the baseline classifier. Hence, the LR classifier has no extra parameters for feature extraction, and the storage cost of weights is moderate. Compared to the CMF, the proposed method trains discriminative confusion set classifiers instead of generative classifiers. Compared to the pre-trained pair discriminators of [6][7] and the neural networks in [8], the proposed method costs only small storage of extra parameters. Our experimental results on online handwritten Japanese characters show that the proposed method largely reduces the recognition errors of non-Kanji characters.

2. System Overview

In our experiments, the baseline recognition system adopts a two-stage classification strategy [10]. After pre-processing and feature extraction of the input pattern, the feature dimensionality is reduced by Fisher linear discriminant analysis (FDA) [11] considering the overall separability of the character classes. On the reduced vector, candidate classes are selected with a coarse classifier according to the Euclidean distance to class means, followed by fine classification with the MQDF [4]. Based on the above system, the third-stage classification is implemented, as illustrated in Fig. 1. A LR classifier with selected subspace features is trained for each confusing set. In testing, the candidate classes output by the baseline classifier are re-classified by the confusing set classifier of the top candidate class.

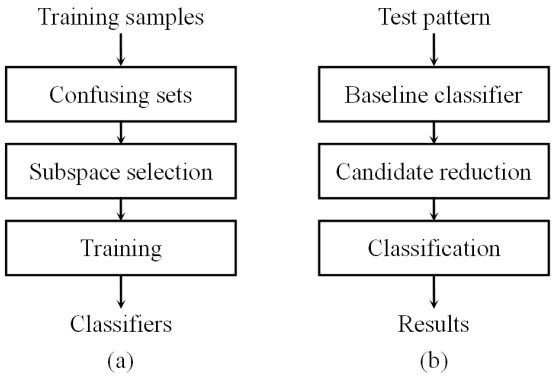

Fig. 1. Training (a) and testing (b) of third-stage classification.

2.1. Confusing Sets Construction

For each class ω_i, i = 1,…, M, we build a confusing set G_i from the training samples by 5-fold cross validation, i.e., rotationally using 4/5 for training the baseline classifier and the remaining 1/5 for validation. G_i is composed of the classes misclassified as ω_i for at least t times, where t is a threshold to control the size of G_i and the risk of misclassification. In the confusion matrix

$$\mathbb{F} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1M} \\ n_{21} & n_{22} & \cdots & n_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ n_{M1} & n_{M2} & \cdots & n_{MM} \end{pmatrix} \quad (1)$$

where n_{ij} denotes the frequency of misclassifying a pattern from ground-truthed class ω_i as class ω_j in cross validation. The confusing set of class ω_i is formed by

$$G_i = \{\omega_c \mid n_{ci} \geq t, c \neq i\}. \quad (2)$$

G_i defines the set that ω_i confuses with. Thus, the classes outside G_i rarely confuse with ω_i when ω_i is the top rank class of the baseline classifier. G_i is empty ($G_i=\phi$) if no class is confused with ω_i according to the above definition. Compared with the approach in [6], which selects all the classes with confidences larger or closer to that of the genuine one, the proposed method selects much fewer confusing classes.

Denoting $\hat{G}_i$ as $\{\omega_i\} \cup G_i$, the task of third-stage classification is to discriminate the classes of each $\hat{G}_i$ when $G_i \neq \phi$. One strategy is to train pairwise classifiers for each $\hat{G}_i$, i = 1,…, M, and determine the recognition result by majority voting or pairwise coupling. Another way is to train a multi-class classifier for each $\hat{G}_i$. Considering the fact that different $\hat{G}_i$ share many confusing classes, to merge the highly overlapping confusing sets can significantly reduce the number of discriminative classifiers.

For two sets $\hat{G}_i$ and $\hat{G}_j$ with $G_i \neq \phi$ and $G_j \neq \phi$, if the ratio

$$\rho = \frac{\left|\hat{G}_i \cap \hat{G}_j\right|}{\left|\hat{G}_i \cup \hat{G}_j\right|} \quad (3)$$

is greater than a threshold, $\hat{G}_i$ and $\hat{G}_j$ are merged, where "$|\cdot|$" denotes the cardinality of a set. Since a smaller set of classes is easier to discriminate, the merging process is relatively conservative. In our experiments, a threshold 0.8 is large enough to avoid over-merge. The confusion sets are merged iteratively until the ratio between each two sets is smaller than the threshold. The confusing sets after merging are

denoted as S_i, $i = 1,\ldots, M'$, where $M' < M$, and a classifier is trained for each S_i for third-stage classification.

2.2. Third-Stage Classification

In testing, the baseline classifier outputs a ranked candidate list L with ω_i as the top rank. If $L \cap \hat{G}_i = \{\omega_i\}$, ω_i is accepted as the recognition result. Otherwise, $L \cap \hat{G}_i$ is re-classified by third-stage classifiers to give the final decision. From the definition of $\hat{G}_i$, we know that with ω_i as the top rank, the classes outside $\hat{G}_i$ rarely confuse with ω_i, so, we only need to discriminate the classes in $L \cap \hat{G}_i$. The classes in $L \cap \hat{G}_i$ are classified by all the classifiers of S_j with $\{S_j \mid L \cap \hat{G}_i \subset S_j, 1 \le j \le M'\}$, and the final decision is determined by majority voting. Because S_i, $i = 1,\ldots, M'$ are merged from $\hat{G}_i$, $i = 1,\ldots, M$, there exists at least one S_j with $L \cap \hat{G}_i \subset S_j$.

3. Third-Stage Classifier Design

In this section, we first formulate the discriminant functions of third-stage classifiers which input the baseline classifier outputs and multiple subspace features, and then describe the subspace selection and weight learning methods. For subspace feature extraction of each class of S_i, $i = 1,\ldots, M'$, K subspace vectors are selected from a common vector set Ψ derived from the baseline classifier.

3.1. Discriminant Function

The third-stage classifiers take the outputs of the baseline classifier and selected subspace features as input. In previous methods of compound distances [5, 6], the output discriminant function of the baseline classifier has been fused with the pair discriminant function for fine classification. The CMF method extracts 1D subspace from the complement subspace of each class, while the LDA-based method learns and stores subspace vectors for confusing pairs in advance. This yields higher classification accuracy than the CMF method but the storage of subspace vectors for a large number of confusing pairs is expensive. To extract discriminant features for our third-stage confusing set classifiers while reducing the storage of feature parameters, we select subspace vectors from a common vector set derived from the baseline classifier.

Denoting $\mathbf{x}$ as the feature vector after dimensionality reduction and $\boldsymbol{\mu}_j$ as the mean vector of class ω_j, the discriminant function of class ω_j with $\omega_j \in S_i$ is formulated as

$$f_{ij}(\mathbf{x}) = a_{ij0} g_j(\mathbf{x}) + \sum_{k=1}^{K} a_{ijk} d_{ijk}(\mathbf{x}) + b_{ij}, \quad (4)$$
$$\omega_j \in S_i,\ i = 1,\ldots, M',$$

where $g_j(\mathbf{x})$ is the output of the baseline classifier for class ω_j, and

$$d_{ijk}(\mathbf{x}) = \left[(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\psi}_{ijk} \right]^2 \quad (5)$$

is the complementary distance on 1D subspace $\boldsymbol{\psi}_{ijk}$ selected from a common vector set, and b_{ij} is the bias term. By Eq. (4), the baseline classifier outputs and the selected complementary distances (subspace features) are combined to discriminate one class from the others in confusing set S_i.

3.2. Subspace Selection

The subspace vectors $\boldsymbol{\psi}_{ijk}$ in Eq. (5) are selected from a common vector set Ψ derived from the baseline classifier. Three basic vector sets of MQDF (our baseline classifier) can provide discriminant information: the axes of the standard coordinate system of the feature space, the class means and the principal eigenvectors of each class. For selecting subspaces of high discriminability, we consider the separability between one class and the rest of each S_i, $i = 1,\ldots, M'$, and select subspaces from the normalized class means: $\Psi = \{\tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k / |\boldsymbol{\mu}_k|, k = 1,\ldots, M\}$. This vector set has a manageable size, and provides good complementary discriminability to the baseline quadradic classifier.

According to Eq. (5), to select subspaces is just to select the projected features. Since the number of candidate features is as large as the class number M, to reduce the time cost of feature selection, we adopt the variable ranking method [12] which considers individual features independently of the others. We select a subspace vector set for each class by considering the separability between the class and the rest in the confusing set. Given a training dataset $\{(\mathrm{x}^n, c^n) \mid n = 1,\ldots, N\}$ (c^n is the class label of x^n), for each class $\omega_j \in S_i$, we rank the subspace vectors $\Psi = \{\boldsymbol{\psi}_{ijk} \mid k = 1,\ldots, K_0\}$ ($K_0 = M$) in decreasing order according to the Fisher criterion for two classes ω_j and $S_i - \{\omega_j\}$:

$$J(\boldsymbol{\psi}_{ijk}) = \frac{\left[(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{ij}^-)^{\mathrm{T}} \boldsymbol{\psi}_{ijk} \right]^2}{\sum_{\mathrm{x}^n \in \omega_j} \left[(\mathbf{x}^n - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\psi}_{ijk} \right]^2 + \sum_{\mathrm{x}^n \in S_i - \{\omega_j\}} \left[(\mathbf{x}^n - \boldsymbol{\mu}_{ij}^-)^{\mathrm{T}} \boldsymbol{\psi}_{ijk} \right]^2}, \quad (6)$$
$$k = 1,\ldots, K_0,$$

where $\boldsymbol{\mu}_{ij}^{-}$ is the mean vector of the training samples belonging to $S_i - \{\omega_j\}$.

After ranking, we select the first K ($K < K_0$) vectors for each $\omega_j \in S_i$. For convenience, we select equal number of subspace vectors for each class. To further speed up feature selection, we first select a reduced set of K_1 ($K<K_1<K_0$) vectors (200 in our experiments) according to the between-class variance (numerator of Eq. (6)), and then select K vectors from the reduced set according to Eq. (6). The variable ranking method based on Fisher criterion has been used effectively in a multi-lingual character recognition system [13], and is known to be closely related to the correlation criterion [12].

3.3. Weight Learning

The classifier for confusing set S_i as in Eq. (4) is a linear classifier. It has K+1 input features for each class: the class output of baseline classifier and K selected subspace features. To learn the weights, we collect training data by 5-fold cross validation, i.e., rotationally 4/5 training samples are used to train the baseline classifier, and the remaining 1/5 samples are used for validation to calculate the baseline classifier outputs and the selected subspace features. The N_i validation samples from the classes $\omega_j \in S_i$ are used to learn the weights Θ_i of the linear classifier for S_i .

We train the linear classifier using the logistic regression (LR) method, which minimizes the multi-class cross-entropy (CE) loss:

$$\min_{\Theta_i} L_{CE} = -\sum_{n=1}^{N_i} \sum_{\omega_j \in S_i} t_j^n \ln p_{ij}^n , \qquad (7)$$
$$i = 1, \ldots, M'$$

where $t_j^n = 1$ if $j=c^n$ (class label of the n-th sample) and 0 otherwise, and

$$p_{ij}^n = \frac{\exp\left(f_{ij}(\mathrm{x}^n)\right)}{\sum_{\omega_c \in S_i} \exp\left(f_{ic}(\mathrm{x}^n)\right)} \qquad (8)$$

By minimizing the CE loss, the weights of Eq. (4) (a_{ijk} and b_{ij}) are iteratively updated by stochastic gradient descent [14].

In gradient learning, the weights in Eq. (4) are initialized to be zero. For good convergence, the input features are divided into two groups: baseline classifier outputs and complementary distances, the features in each group are rescaled to zero mean and standard deviation one with the global mean and variance of these features on the samples of the classes in each S_i.

Since the input features of confusing set classifier is class-specific, the total number of features for each sample of class ω_j is $\sum_{S_i \supset \{\omega_j\}} |S_i| \times (K+1)$, where $|S_i|$ denotes the class number of S_i. It is hard to store the features of all samples for offline training. We instead generate the sample features only during training. By 5-fold cross validation, five baseline classifiers, one for each rotational 4/5 dataset, have been trained. And for each confusing set, subspace feature selection has been performed. Then in LR training, the class-specific features for a sample in a 1/5 validation set is calculated from its corresponding baseline classifier in real time whenever it is used to update the weight parameters in an iteration.

4. Experimental Results

We evaluated the performance of the proposed method on the TUAT HANDS databases, Kuchibue and Nakayosi, of online handwritten Japanese characters [15]. To compare with our previous work [10], we experimented with 3,345 classes (2,965 JIS level-1 Kanji characters and 380 non-Kanji characters) as well as 2,965 Kanji classes only, using the samples of Nakayosi for training classifiers and the samples of Kuchibue for testing.

From each character pattern, we extract 512D feature of local direction histogram, with the trajectory normalized by the moment normalization method in original direction [10]. The 512D is reduced to 160D by LDA. The baseline classifier is the MQDF, with 50 principal eigenvectors for each class. The parameters of MQDF include the class means, which are used to select candidate classes in the first stage according to Euclidean distance.

With the baseline recognition system, we observed a big gap between the accuracies of 3,345-class recognition (90.51%) and Kanji recognition (97.90%), because there are many confusing classes among the 380 non-Kanji characters and between non-Kanji and Kanji characters. For third-stage classification, we consider only the top five rank classes of the baseline classifier with the accumulative accuracy 99.02% for 3,345-class recognition and 99.75% for Kanji recognition.

For 3,345 classes, the recognition accuracies for different threshold t with varying number of subspace features are shown in Fig. 2. Table 1 lists the recognition accuracies, increased storage and time cost (on an Intel Quad Core 2.83GHz CPU 4 GB-RAM PC) for different threshold t with K=100 subspace features, and Table 2 lists the number of third-stage classifiers (M') and numbers of confusing classes.

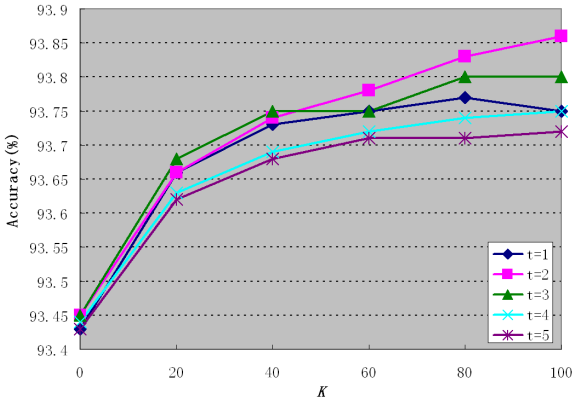

Fig. 2. Test accuracies of 3,345-class recognition with varying number of subspace features.

Table 1. Test accuracies of all samples, Kanji, non-Kanji, increased storage, training time and average testing time of 3,345-class recognition with *K*=100. For MQDF, storage denotes the dictionary size.

	MQDF	*t*=1	*t*=2	*t*=3	*t*=4	*t*=5
All (%)	90.51	93.75	93.86	93.80	93.75	93.72
Kanji (%)	96.09	96.43	96.41	96.37	96.34	96.31
Non-Kanji (%)	85.70	91.36	91.58	91.51	91.44	91.41
Storage (MB)	105.11	7.65	3.03	2.03	1.61	1.34
Train time (h)	1.14	81.96	35.52	25.18	20.46	17.22
Test time (ms)	3.75	4.33	4.32	4.12	4.04	4.00

Table 2. Classifier number and maximum, average class number of S_i for 3,345-class recognition.

	t=1	*t*=2	*t*=3	*t*=4	*t*=5
#classifier	2493	1275	881	708	615
Maximum	36	22	17	13	12
Average	5.15	3.88	3.66	3.54	3.35

Fig. 2 shows that third-stage classification can effectively reduce confusion errors. Compared to the MQDF (90.51%), even without subspace features (*K*=0), the recognition accuracies are greatly improved due to discriminative learning and class set reduction. With increasing number of subspace features, the classification accuracy further increases.

From Table 1 and Table 2 we can see that by increasing the threshold *t* to reduce the number of confusing classes, the extra storage and time cost are reduced, especially the training complexity is greatly alleviated, while the recognition accuracies are comparable. In contrast to the baseline MQDF, third-stage classification improves the recognition accuracies with only slightly increased storage and recognition time. By comparing the recognition accuracies for all test samples, Kanji and non-Kanji samples of the MQDF, we know that the confusion is mainly caused by non-Kanji characters. By third-stage classification, the recognition accuracy of Kanji characters is improved slightly, while that for non-Kanji characters is improved prominently. This justifies that third-stage classification effectively reduces the confusion caused by non-Kanji characters.

Some samples which are misclassified by MQDF while corrected by third-stage classification are shown in Fig. 3. We can see that third-stage classification can discriminate characters with slight shape difference.

Note that the patterns of some non-Kanji classes such as letters ‘O’, ‘o’ and numeral ‘0’ have identical shape after normalization, and it is almost impossible to separate them at character level. Nevertheless, the proposed third-stage classification method still improves the overall recognition accuracy remarkably.

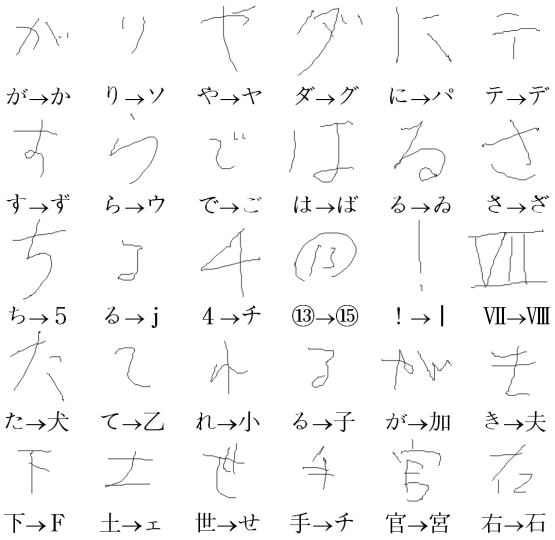

Fig. 3. Samples misrecognized by MQDF, but corrected by third-stage classification with *t*=2 and *K*=100.

We also performed experiments of recognizing 2,965 JIS level-1 Kanji characters only. Since the Kanji classes have relatively fewer training samples compared to the non-Kanji characters [15], to alleviate overfitting, we reduce the model complexity by sharing weights of the subspace features for each class in confusing set S_i. With 50 subspace features, the recognition results are listed in Table 3. It is shown that the accuracies of Kanji recognition is hardly improved by third-stage classification compared to the baseline MQDF. This is because the Kanji characters are less confused on one hand and the number of training samples is small for third-stage classification

on another hand. We conjecture that to separate similar Kanji characters, subspace features with more discriminability should be incorporated.

Table 3. Test accuracies (%) and increased storage (MB) of 2,965-class Kanji recognition with *K*=50.

	MQDF	*t*=1	*t*=2	*t*=3	*t*=4	*t*=5
Accuracy	97.90	97.77	97.88	97.92	97.94	97.95
Storage	93.20	2.12	0.70	0.46	0.34	0.28

5. Conclusion

In this paper, we propose an error reduction method for online handwritten Japanese character recognition by discriminating confusing classes using multi-class logistic regression (LR) classifiers. With the subspace features of LR classifiers selected from existing vectors of the baseline MQDF classifier, the storage space increases only slightly. The experiments on TUAT HANDS databases demonstrate that the proposed method can effectively reduce the confusion errors caused by non-Kanji characters, and the overall accuracy is improved remarkably. It is our future work to consider subspace features with more discriminability for further higher accuracy.

Acknowledgements

This work is supported by the R&D fund for "Development of Pen & Paper-Based User Interaction" of Japan Science and Technology Agency and the National Natural Science Foundation of China (NSFC) grant no.60825301. The authors thank Dr. Bilan Zhu for sincere help.

References

[1] C.L. Liu, S. Jaeger, M. Nakagawa, Online recognition of Chinese characters: the state-of-the-art, *IEEE Trans. Pattern Anal. Mach. Intell.* 26(2) (2004) 198-213.

[2] C.L. Liu, H. Fujisawa, Classification and learning methods for character recognition: Advances and remaining problems, In: *Machine Learning in Document Analysis and Recognition*, S. Marinai and H. Fujisawa (Eds.), Springer, 2008, pp.139-161.

[3] A.F.R. Rahman, M.C. Fairhurst, Multiple classifier decision combination strategies for character recognition: A review, *Int. J. Document Analysis and Recognition* 5(4) (2003) 166-194.

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

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

[6] T.F. Gao, C.L. Liu, High accuracy handwritten Chinese character recognition using LDA-based compound distances, *Pattern Recognition* 41(11) (2008) 3442-3451.

[7] K.C. Leung, C.H. Leung, Recognition of handwritten Chinese characters by critical region analysis, *Pattern Recognition* 43(3) (2010) 949-961.

[8] Y. Kimura, T. Wakahara, A. Tomono, Combination of statistical and neural classifiers for a high-accuracy recognition of large character sets, *Trans. IEICE Japan* J83-D-II (10) (2000) 1986-1994.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning,* Springer, 2006.

[10] C.L. Liu, X.D. Zhou, Online Japanese character recognition using trajectory-based normalization and direction feature extraction, *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition*, La Baule, France, 2006, pp. 217-222.

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

[12] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *J. Machine Learning Research* 3 (2003) 1157-1182.

[13] H.S. Park, H.H. Song, S.W. Lee, A self-organizing hierarchical classifier for multi-lingual large-set oriental character recognition, *Int. J. Pattern Recognition and Artificial Intelligence* 12(1) (1998) 191-208.

[14] H. Robbins, S. Monro, A stochastic approximation method, *Ann. Math. Stat.* 22 (1951) 400–407.

[15] K. Matsumoto, T. Fukushima, M. Nakagawa, Collection and analysis of on-line handwritten Japanese character patterns, *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, Seattle, WA, 2001, pp. 496-500.