Historical Hand-Written String Recognition by Non-linear Discriminant Analysis using Kernel Feature Selection

Ryo Inoue, Hidehisa Nakayama, and Nei Kato
Graduate School of Information Sciences, Tohoku University, Sendai, Japan
r-ino@it.ecei.tohoku.ac.jp

Abstract

In this paper, we propose a method to compose a classifier by non-linear discriminant analysis using kernel method combined with kernel feature selection for holistic recognition of historical hand-written string. Through experiments using historical hand-written string database HCD2, we show that our approach can obtain high recognition accuracy comparable to that of individual character recognition.

1. Introduction

Currently, most hand-written character recognitions have mainly focused on individual character recognition. As a result, the recognition for hand-written character database can achieve a very high acuracy, for example, the accuracy for $3,036$ characters of Hiragana and Chinese character in ETL9B is reported to be more than 99%. While the accuracy for a hand-written character string is lower than hand-written character with about 90% [4]. As we can see, the recognition accuracy for the historical hand-written string is lower than that for modern Japanese hand-written string.

Most string recognition approaches proceed by segmenting string images into individual characters which are recognized separately, and the string is recognized as the composition of recognized parts. However, this induces a serious problem since such series method may lead to heap up errors at each module. To avoid such problem holistic recognition approach appears to be an attractive solution [5]. Holistic recognition method uses the strategy similar to individual character recognition by treating the whole string of Japanese document image as a single oblong.

In holistic recognition, since the string is one unit, it is necessary to extract high-dimensional feature because the string has several characters. However, it is almost impossible to collect a huge amount of string images as training data. Hence, the number of training samples become relatively small compared to the data dimensionality, which makes it difficult to determine a classifier parameter accurately. For this reason, when determining the classifier parameter, it is necessary to select a significant feature. Historical hand-written string image consists of running-hand and the sample size in each category is too small. Since the ratio of deformation characters is higher than in modern Japanese string, when determining a classifier parameter, an outlier data impairs the performance of a classifier. Thus, in order to improve the accuracy, the sample selection is more preferable compared to individual character recognition.

In this paper, we propose a method to compose a classifier by non-linear discriminant analysis using kernel method combined with kernel feature selection. We applied Kernel Discriminant Analysis (KDA) to the holistic recognition of historical hand-written string. Our experiments with historical hand-written string database HCD2 [7] clearly show the effectiveness of the proposed method which obtains high accuracy equivalent to individual character recognition.

This paper is organized as follows. Section 2 explains the traditional holistic recognition system using canonical discriminant analysis. Section 3 reveals the problem of traditional system, and proposes a new holistic recognition system using kernel discriminant analysis. In Section 4, we show the efficiency of our system through experiments. Section 5 concludes the paper.

2. Holistic recognition system

Fig.1 shows the basic composition of holistic recognition system for the hand-written string used in this paper.

2.1. Image pre-processing

After applying smoothing and noise erasing to a binary input image, we normalize it non-linearly to expand it to the entire area. Non-linear shape normalization is performed by three step process; determination of the line density by setting the image characteristic value, derivation of transform function, and reverse-mapping. Several methods have been proposed according to the definition of the characteristic value to determine the line density.

In linear shape normalization (LSN) , a constant is used as the characteristic value. In this paper, we adopt a non-

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

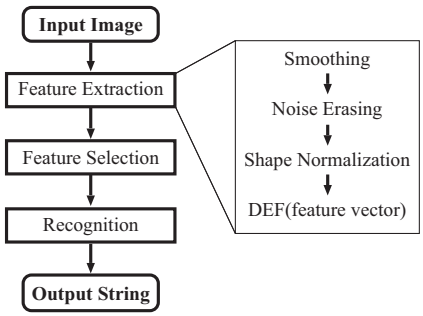

Figure 1. Basic Holistic Recognition System

linear shape normalization (NSN) using the average line interval between the stroke part and the background part and using the blurring line density smoothed by Gaussian filter [3].

2.2. Feature Extraction

We create the contour image that consists of character outlines by image processing. Then we calculate Directional Element Feature [6] from this contour image. In this paper, original feature is extracted as $1,540$ dimensions.

2.3. Feature Selection

Compared to individual character image, the number of string image samples that can be gathered is extremely small. As a consequence, the number of sample is very small compared to the feature dimensionality, which makes it difficult to determine a classifier parameter accurately. To avoid such problem, we perform feature selection on the original feature and reduce the dimensionality from d_0 to d in advance. In this paper, we adopt Forward Stepwise Selection (FSS) as a feature selection method [2] and F-test, significance level of which is 5%, is used to select the significant elements.

2.4. Recognition

As mentioned in 2.3, it is difficult to gather sufficient data of string for each category. Therefore, in this paper, we use Canonical Discriminant Analysis (CDA) for holistic recognition. CDA has a comparatively high generalization ability even when the number of samples in each category is small.

First, between-class and within-class scatter matrix $\boldsymbol{S_B}$, $\boldsymbol{S_W}$ are calculated using the training samples of all categories. Next, the eigenvalue matrix $\boldsymbol{\Lambda}$ and the eigenvector matrix $\boldsymbol{A}$ are obtained by solving the generalized eigenvalue problem as follows:

$$\boldsymbol{S_B A} = \boldsymbol{S_W A \Lambda} \tag{1}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_d)$, λ_i is eigenvalue ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$) and $\boldsymbol{A} = [\boldsymbol{a_1 a_i} \cdots \boldsymbol{a_d}]$, $\boldsymbol{a_i}$ is eigenvector corresponds to eigenvalue λ_i. Then, vector $\boldsymbol{y}$ made by the feature selection in 2.3 is projected using this eigenvector, and the canonical value is calculated. When the number of category is assumed to be c, each data can be expressed by only $(c-1)$ canonical values as follows:

$$z_i = \boldsymbol{a_i}^T \boldsymbol{y} \quad (i = 1, 2, \cdots, c-1) \tag{2}$$

We define the recognition rule as follows: a test data belong to the category whose Euclidean distance to the mean of the training data is the smallest in canonical vector space.

3. Improvement of classifier with kernel method and kernel feature selection

In this paper, we use a holistic recognition method where the feature is extracted directly without segmenting the string image. Since the string is treated as one unit, it is difficult to absorb the differences between each character size. Therefore, the proportion of the pixel is low and many blank parts exist in the image. In other word, there are fewer feature parts that are significant for classifier, compared to the individual character recognition. As one of the techniques to avoid such problem, classifier using feature selection seems to be a good solution for holistic recognition, described in 2.3. However, this technique is not definitely versatile and it is hard to say that it can represent the detailed feature of a character. In order to improve the recognition accuracy of holistic recognition, it is necessary to use the classifier that is capable of using the original feature vector without omitting the significant parts of character.

A classifier using Kernel method appears to be a promising technique which is able to use the feature vector without reducing the significant parts. By using Kernel method, we can reduce the computation cost from the order of original feature dimensionality to that of the number of training sample. The Kernel method is convenient when the number of training sample is small and the feature dimensionality is very high, such as the holistic recognition. For this reason, in this paper, we adopt a classifier that uses Kernel Discriminant Analysis (KDA) [1].

3.1. Kernel method

In the kernel method, non-linear discrimination is performed in original space by non-linear mapping of the original feature vector from the original space $\mathcal{R}^d$ to the high-dimensional feature space $\mathcal{F}$, and linear discriminating on $\mathcal{F}$. However, in general, the dimensionality of F is very high compared to $\mathcal{R}^d$, or possibly infinite. Therefore, it is difficult to calculate the projection of feature vector positively. Consequently, the kernel trick is well-known in the kernel method. Let $\boldsymbol{x}, \boldsymbol{y}$ be the original feature vector and Φ be non-linear mapping, dot product of Φ is expressed as follows:

$$\Phi(\boldsymbol{x})^T \Phi(\boldsymbol{y}) = K(\boldsymbol{x}, \boldsymbol{y}) \tag{3}$$

where K is kernel function. That is, the dot product on $\mathcal{F}$ is equivalent to the kernel function on $\mathcal{R}^d$. By using this scheme, the linear classifier composed of the dot product on $\mathcal{R}^d$ can be applied as a non-linear classifier. In other word, this non-linear classifier can be calculated using the kernel function of the original feature vector. This scheme is called a kernel trick. Many functions have been proposed as kernel functions. In this paper, Sigmoid kernel is used.

$$K(\boldsymbol{x}, \boldsymbol{y}) = 1 + tanh(\boldsymbol{x}^T \boldsymbol{y}) \tag{4}$$

3.2. Kernel discriminant analysis (KDA)

In the KDA, we consider about the linear discriminant mapping $\boldsymbol{a_j}$ $(i = 1, \cdots, p)$ on $\mathcal{F}$, and the canonical discriminant vector $\boldsymbol{z}$ is calculated using the set of optimal discriminant matrix A as follows:

$$\boldsymbol{z} = \boldsymbol{A}^T \Phi(\boldsymbol{x}) \quad \boldsymbol{A} = \left[\boldsymbol{a_1} \ \cdots \ \boldsymbol{a_p}\right] \tag{5}$$

where $1 \leq p < c$, c is the number of category. Let $\boldsymbol{x_i}$ $(i = 1, \cdots, N)$ be the training samples, the linear discriminant mapping on $\mathcal{F}$ is expressed using linear combinations with coefficients u_{ij} such that:

$$\boldsymbol{a_j} = \sum_{i=1}^{N} u_{ij} \Phi(\boldsymbol{x_i}), \quad (j = 1, \cdots, p) \tag{6}$$

This is substituted to equation (5) and calculated as follows:

$$\begin{aligned} \boldsymbol{z} &= \left[\textstyle\sum_i u_{i1} \Phi(\boldsymbol{x_i})^T \Phi(\boldsymbol{x}) \ \cdots \ \sum_i u_{ip} \Phi(\boldsymbol{x_i})^T \Phi(\boldsymbol{x})\right] \\ &= \left[\textstyle\sum_i u_{i1} K(\boldsymbol{x_i}, \boldsymbol{x}) \ \cdots \ \sum_i u_{ip} K(\boldsymbol{x_i}, \boldsymbol{x})\right] \\ &= \boldsymbol{U}^T \vec{K}(\boldsymbol{x}), \quad \boldsymbol{U} = \left[\boldsymbol{u_1} \ \cdots \ \boldsymbol{u_p}\right] \end{aligned} \tag{7}$$

where $\vec{K}(\boldsymbol{x}) = \left[K(\boldsymbol{x_1}, \boldsymbol{x}) \ \cdots \ K(\boldsymbol{x_N}, \boldsymbol{x})\right]^T$ is kernel feature vector. From this expression, between-class and within-class scatter matrix $\boldsymbol{S_B^K}$, $\boldsymbol{S_W^K}$ concerning the kernel feature vector are obtained. The eigenvalue matrix $\boldsymbol{\Lambda_U}$ and the eigenvector matrix $\boldsymbol{U}$ are obtained by solving the generalized eigenvalue problem as well as CDA as follows:

$$\boldsymbol{S_B^K U} = \boldsymbol{S_W^K U \Lambda_U} \tag{8}$$

In short, KDA is equivalent to solving CDA concerning the kernel feature vector.

Similar to 2.4, we define the recognition rule as follows: a test data belong to the category whose Euclidean distance to the mean of the training data is the smallest in canonical vector space.

3.3. Kernel feature selection

In KDA, the classifier parameter is calculated from the kernel feature vector. In 2.3, we applied dimensionality reduction method FSS to accurately calculate the classifier parameter on insufficient training samples. When the kernel method is used to transform the original feature vector to the kernel feature vector which is called "kernel mapping", the drawback of high dimensionality of the original feature vector, is alleviated by calculating the kernel function. Therefore, in 3.2 we do not use the feature selection of the original feature. On the other hand, regarding to the kernel feature vector, the drawback of the high dimensionality has been concerned. Therefore, after kernel mapping, the kernel feature selection is performd by FSS based on Wilks's Λ statistic, similar to 2.3.

Next, we consider about an effect of this kernel feature selection. The kernel feature vector is calculated by the kernel function of the original feature vector with all training samples. That is, each element of the kernel feature vector represents the correlation regarding each training sample. Therefore, besides having an effect in dimensionality reduction, kernel feature selection also has an effect in sample selection. As a result, if the kernel feature selection is performed, we can decrease the influence of an outlier, and improve the generalization ability.

From the above-mentioned, the proposed holistic recognition system of the string is illustrated in Fig.2.

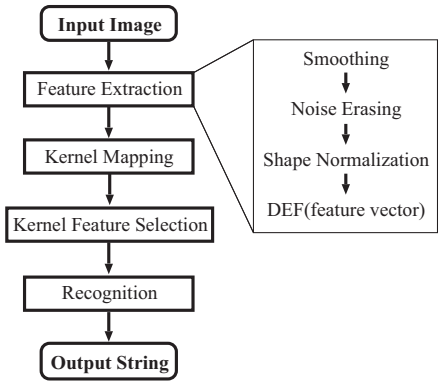

Figure 2. Proposed Holistic Recognition System

4. Experimental evaluation

For experimental evaluation, total 112 samples of the 5 categories in historical hand-written string database HCD2 made in HCR(Historical Character Recognition) project [7] are used as an evaluation for the recognition experiment.

4.1. Experimental method

Here, we refer to the holistic recognition system described in Section 2 as a traditional method, and the improved method presented in Section 3 as a proposed method. This experiment is conducted to compare the recognition rate of the traditional method with the proposed method. For shape normalization in the image pre-

processing, we used two techniques (LSN, NSN) described in 2.1. The CDA recognition rate for different value of selected dimensionality of original feature is obtained by using the method described in 2.3, and the KDA rate for dimensionality of kernel feature is obtained by using the method described in 3.3. The recognition rate is evaluated using the Leave-One-Out method which divides all 112 samples into 111 training samples and 1 test sample. Then experiments performed by 112 times.

4.2. Experimental results

Fig.3, 4 shows the experimental results with various normalized images. The results are summarized as Table 1,2.

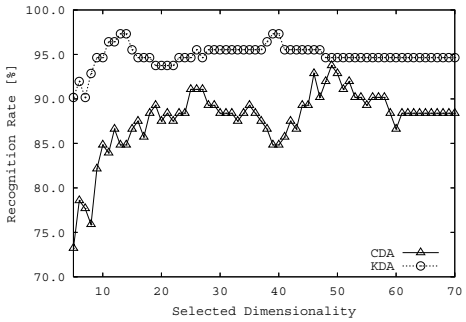

Figure 3. Experimental Results on Historical String Images using LSN

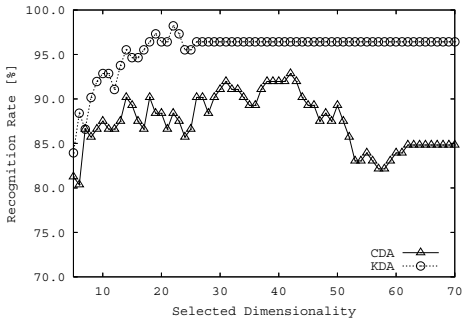

Figure 4. Experimental Results on Historical String Images using NSN

We consider about linear classifier using feature selection and non-linear classifier using kernel feature selection using result. From Fig.3, 4, we can clearly see that KDA performs better than CDA as it gives higher recognition rate in all the selected dimensionality. We can consider that the proposed system can reduce insignificant element because F-test performs well in kernel feature selection, compared with the system without kernel feature selection. The effect appears remarkably in KDA than CDA, since using KDA, we can achieve higher accuracy in lower dimensionality than CDA, it is possible to reduce the calculation cost. Therefore, we can say that the proposed system is more effective than the traditional system.

Table 1. Results of Maximum Recognition Rate and Difference

	CDA	KDA	Diff.
LSN	93.8%	97.3%	3.5%
NSN	92.9%	98.2%	5.3%

Table 2. Results of Satulation Number of Dimension

	CDA	KDA
LSN	$61-111\ (88.4\%)$	$48-111\ (94.6\%)$
NSN	$62-111\ (84.8\%)$	$26-111\ (96.4\%)$

5. Conclusion

In this paper, we investigated the classifier to the holistic recognition system for historical hand-written string. In holistic recognition, it is necessary to extract high-dimensional feature. When composing the classifier that handles high-dimensional feature with small sample, it is indispensable to perform feature selection. However, the feature selection to the original feature often gives poor recognition accuracy. By adopting kernel method which does not perform feature selection to the original feature, we could improve the traditional system and obtain high performance. In addition, we showed that by performing kernel feature selection, we can achieve dimensionality reduction as well as sample selection which can improve accuracy.

In our experiments on historical hand-written string database HCD2, the proposed system can obtain recognition rate of 98.2% when evaluated using Leave-One-Out method. We showed the effectiveness of the proposed system.

References

[1] G. Baudat et al. Generalized discrimnant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.

[2] K. Fukunaga. *Statistical Pattern Recognition(2nd ed.)*. Academic Press, NY, 1990.

[3] T. Horiuchi et al. Two-dimentional extension of nonlinear normalization method using line density. *Trans. IEICE Japan*, J80–D–II(6):1600–1607, June 1997.

[4] C. Liu et al. Lexicon-driven segmentation and recognition of handwritten character strings for japanese address reading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(11):1425–1437, Nov. 2002.

[5] S. Madhvanath, E. Kleinberg, and V. Govindaraju. Holistic verification of handwritten phrases. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(2):149–164, Feb. 2001.

[6] N. Sun, M. Abe, and Y. Nemoto. A handwritten character recognition system by using improved directional element feature and subspace method. *Trans. IEICE Japan*, J78–D–II(6):922–930, June 1995.

[7] S. Yamada and M. Shibayama. A stydy of character recognition for historical documents. *IPS Magazine Japan*, 43(9):950–955, Sept. 2002.