Precise Recognition of Blurred Chinese Characters by Considering Change in Distribution

Shin'ichiro Omachi, Fang Sun, and Hirotomo Aso

Department of Communication Engineering,
Faculty of Engineering, Tohoku University
Aza-Aoba, Aramaki, Aoba-ku, Sendai-shi, 980-77 Japan

Abstract

In this paper, a new algorithm for Chinese character recognition is presented. This method has an especial effect on character images those are not clear. Comparing with a clear character image, a blurred character image always has a big change in shape that makes the recognition more difficult. In order to recognize blurred characters precisely, the concept of degree of blur is employed. We also investigate how the distribution of feature vectors changes if a part of character image is crashed. Since the degree of blur is calculated from each sub-area of character image, the condition of the image can be described clearly. The distribution can be reconstructed to respond to the change expressed by the degree of blur. The method proposed in this paper can recognize blurred characters precisely by modifying the Mahalanobis distance function according to change in distribution. The effectiveness of the new method is shown by experiments with blurred character images.

1 Introduction

Automatic Chinese character (KANJI) recognition system is necessary to convert large volumes of documents into character codes readable by computers. Since Chinese characters have complex structures, if the character images are copied or transfered with facsimile, they are almost impossible to be recognized by ordinary methods because of damage to shape of images. It is necessary to develop a new optical character reader (OCR) to recognize those characters of poor quality.

In most traditional methods, distribution of feature vectors is estimated for each category using training patterns. When an unknown input pattern is given, a category with the maximum probability for the pattern is selected as the candidate. But, if the input character image is blurred, the shape of the image is quite different from the original one and the distribution of feature vectors also changes a lot. In this case, probability density function is considered to be necessary to respond to the change.

The authors have proposed *degree of blur*, that can clearly describe condition of each sub-area of a character image[4]. In this paper, we use the concept of degree of blur and investigate how the distribution of feature vectors changes if a part of character image is crashed. By modifying the Mahalanobis distance function according to the change in distribution, a new method that can recognize blurred Chinese characters precisely is proposed.

2 Degree of Blur

2.1 Definition

The Directional Element Feature[7] is used as the feature vector here, and it is calculated as follows. As shown in Fig. 1, input image is normalized to 64×64dots and thinned. Then it is divided into 49 sub-areas of 16×16dots where each sub-area overlaps 8 dots of the adjacent sub-area. (For example, meshed area of Fig. 1d.) For each sub-area, a four-dimensional vector is defined to represent the quantities of the four orientations which are vertical, horizontal, and two oblique lines slanted at $\pm 45°$. Thus the total vector for one character has 196 ($= 49 \times 4$) dimensions.

Thinning is a repeating process of erasing a black pixel from boundaries of black pixels of a character image. By scanning ten pixels around each black pixel, a character image is finally erased to one-pixel width[1]. Figs. 2a

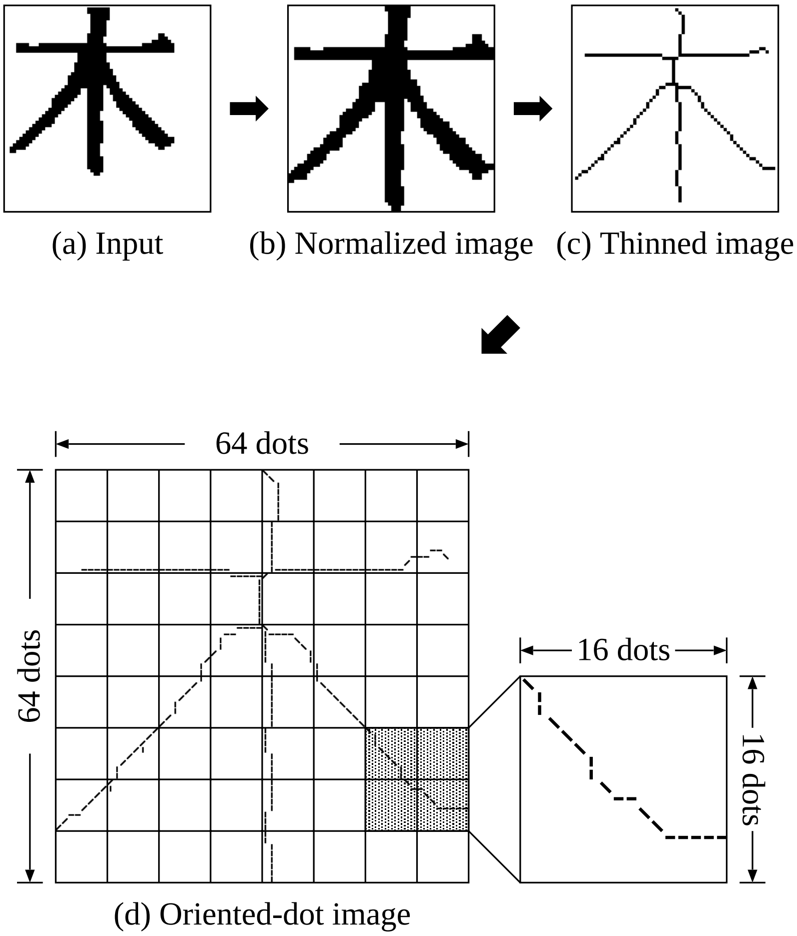


Figure 1: Directional Element Feature.

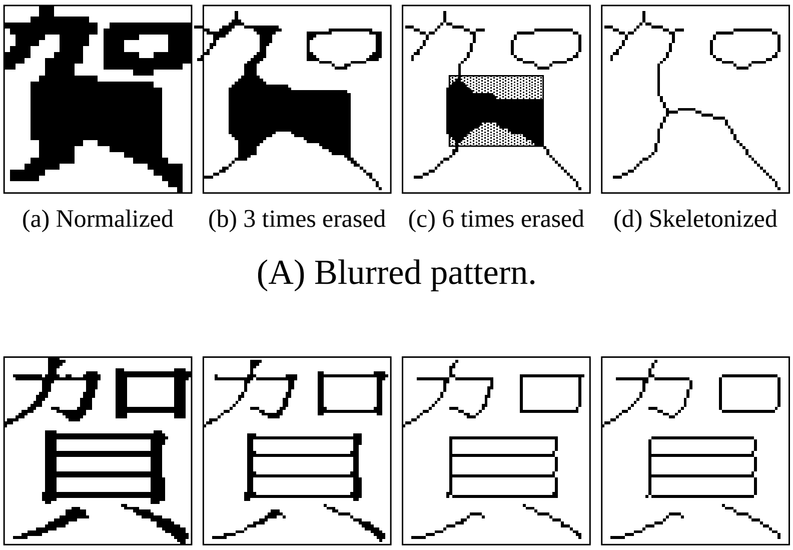
and 2e are examples of normalized blurred image and clear image. Erasing process of these images are shown by Figs. 2b∼2d and Figs. 2f∼2h, respectively. In order to get completely skeletonized image of Fig. 2d, we have to erase pixels on boundaries 14 times, while Fig. 2h requires only 4 times.

If the repeating time in thinning process is limited to the number that clear image is completely skeletonized, obviously it will not be enough for the blurred one. Because our result of pre-experiment has shown that six times is the most suitable number for skeletonizing a clear image, we call an image is blurred if its line width is more than one-pixel after being thinned six times. Usually, an image includes both clear parts and blurred parts. With the maximum number of repeating time, blurred parts will not be thinned to one-pixel width, for example the meshed area of Fig. 2c. For every 49 sub-areas the number of black pixels that are not located at boundaries is counted, and the value of $\lfloor(\text{Number of black pixels})/32\rfloor$ is defined as degree of blur. Since each sub-area has $16 \times 16 = 256$ dots, the value of degree of blur is zero or positive integer not greater than 8. The greater the value is, the more terrible the blur is.

2.2 Characteristic

If a character image has been damaged by copy or facsimile, the appearance of the image is always blurred. Degree of blur is defined to describe the state of each sub-area of an image. We investigate how the standard deviation of each element of feature vector changes according to degree of blur. For this purpose, printed single font 2,965 kinds of Chinese characters of ten sizes (from 6 point to 22 point) are used to carry out a pre-experiment. All these sample patterns are scanned by an optical scanner and transformed to feature vectors.

Let x_{ij}^k be the ith element of a feature vector of jth sample pattern of category k, and b_{ij}^k be the degree of blur corresponding to x_{ij}^k. First, the mean value $\bar{x}_i^k$ and the standard deviation σ_i^k are calculated from the sub-areas



(a) Normalized (b) 3 times erased (c) 6 times erased (d) Skeletonized

(A) Blurred pattern.

(e) Normalized (f) 1 time erased (g) 2 times erased (h) Skeletonized

(B) Clear pattern.

Figure 2: Thinning process (character “賀”).

whose degree of blur is zero. That is,

$$\bar{x}_i^k = \frac{1}{|J_i^k|} \sum_{j \in J_i^k} x_{ij}^k, \tag{1}$$

$$\sigma_i^k = \sqrt{\frac{1}{|J_i^k|} \sum_{j \in J_i^k} (x_{ij}^k - \bar{x}_i^k)^2} \tag{2}$$

where the set $J_i^k = \{j | b_{ij}^k = 0\}$ and $|J_i^k|$ is the number of elements in the set J_i^k. Then each value x_{ij}^k is normalized as $\hat{x}_{ij}^k = (x_{ij}^k - \bar{x}_i^k)/\sigma_i^k$.

Next, standard deviation r_b is calculated with all the value $\hat{x}_{ij}^k$ whose degree of blur is b $(0 \leq b \leq 8)$. That is,

$$r_b = \sqrt{\frac{1}{|D_b|} \sum_{(i,j,k) \in D_b} (\hat{x}_{ij}^k - m_b)^2} \tag{3}$$

where the set $D_b = \{(i,j,k) | b_{ij}^k = b\}$ and $m_b = \frac{1}{|D_b|} \sum_{(i,j,k) \in D_b} \hat{x}_{ij}^k$. In other words, the ratio of standard deviation of the areas with b degrees of blur to the areas with 0 degree is calculated.

The result is shown in Fig. 3. Horizontal axis displays the degree of blur and vertical axis shows the ratio of standard deviations. The figure shows that the larger the degree of blur is, the larger the ratio of standard deviation is. This means that the distribution of the feature vectors is changed with degree of blur.

3 New Evaluation Function Considering Change in Distribution

First we consider the Mahalanobis distance. Let n be the dimension of the feature vector, in the case of the Directional Element Feature, $n = 196$. Let $\boldsymbol{\mu}$ and Σ be the mean vector and the n_by_n covariance matrix, respectively. The squared Mahalanobis distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}$ is defined as

$$d_M^2 = (\boldsymbol{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{4}$$

$$= \sum_{i=1}^{n} \frac{1}{\lambda_i} ((\boldsymbol{x} - \boldsymbol{\mu})^t \boldsymbol{\phi}_i)^2. \tag{5}$$

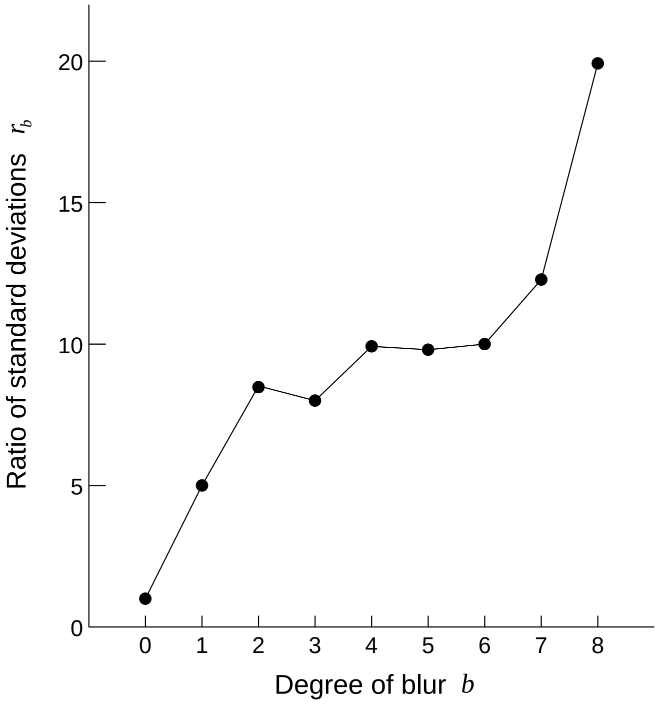


Figure 3: Relationship between degree of blur and ratio of standard deviations.

Here, λ_i is the ith eigenvalue of Σ sorted by descending order, and $\boldsymbol{\phi}_i$ is the eigenvector that corresponds to λ_i. The squared Mahalanobis distance is abbreviated as Mahalanobis distance below.

Next, we investigate how distribution changes and how the evaluation function should be changed if an image is blurred. Because degree of blur is a definite measure that shows how much the distribution of each 49 sub-area has been blurred, by changing the distribution of the blurred part of image according to degree of blur, the distribution can be reconstructed to respond to standard patterns well. For simplicity, we consider the case of two-dimensional normal distribution. As shown in Fig. 4a, let $\boldsymbol{\phi_1}$ be the eigenvector and λ_1 be the eigenvalue that correspond to the first principal component. If the area corresponding to $\boldsymbol{e}_1$-axis is blurred and the area corresponding to $\boldsymbol{e}_2$-axis is clear, the standard deviation on $\boldsymbol{e}_1$-axis will become r_b times larger, and that on $\boldsymbol{e}_2$-axis does not change. Here r_b is the value of vertical axis of Fig. 3 that responds to degree of blur. In this case, the distribution is considered to change as Fig. 4b. Obviously, since the standard deviation on $\boldsymbol{e}_1$-axis becomes r_b times larger, the distribution of Fig. 4b is not normal. To compensate the larger deviation, the term of Eq. 5 corresponding to $\boldsymbol{e}_1$-axis should be multiplied by $1/r_b^2$.

In the case of n dimensions, let $b(i)$ be the degree of blur of a sub-area corresponding to $\boldsymbol{e}_i$-axis. An array $K = [k_{ij}]$ is defined as

$$k_{ij} = \begin{cases} 1/r_{b(i)} & (i = j) \\ 0 & (i \neq j). \end{cases} \tag{6}$$

Let λ_i be the ith eigenvalue of Σ sorted by descending order, and $\boldsymbol{\phi}_i$ be the eigenvector that corresponds to λ_i. The following evaluation function is thought to be more valid for recognition.

$$\hat{d_M^2} = \sum_{i=1}^{n} \frac{1}{\lambda_i}((\boldsymbol{x} - \boldsymbol{\mu})^t K \boldsymbol{\phi}_i)^2. \tag{7}$$

But it is known that the Mahalanobis distance has disadvantages such as the large computation time and bad influence caused by limited samples. To resolve these problems, *Simplified Mahalanobis distance* (SMD) that has the same statistical properties as the Mahalanobis distance is proposed[6]. The effectiveness of SMD has been shown with ETL9B[5], which is the largest database of handwritten Chinese characters in Japan. As it is known that handwritten characters always contain some blurred or crashed parts, SMD can be considered as an effective

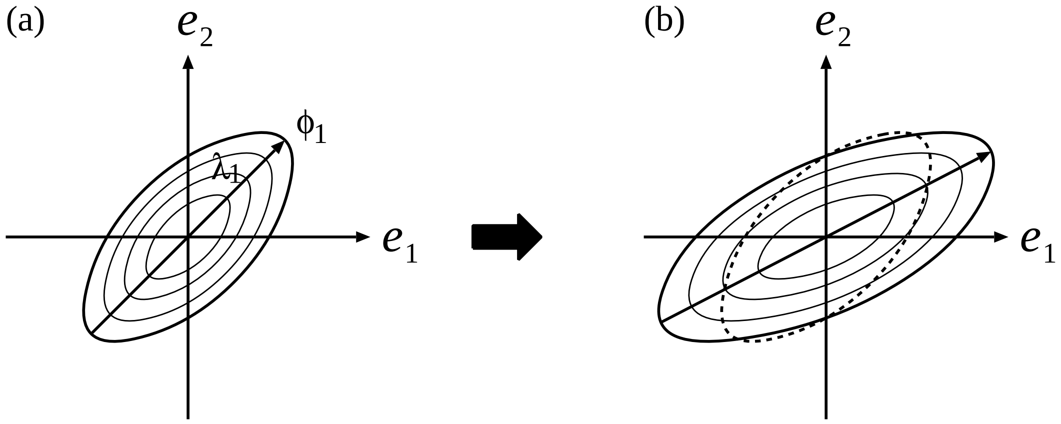


Figure 4: Change in distribution.

evaluation function for blurred character recognition. SMD is given as

$$d_S^2 = \sum_{i=1}^{m} \frac{1}{\lambda_i}((\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\phi}_i)^2 + \sum_{i=m+1}^{n} \frac{1}{\lambda}((\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\phi}_i)^2 \tag{8}$$

$$= \sum_{i=1}^{m} \frac{1}{\lambda_i}((\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\phi}_i)^2 + \frac{1}{\lambda}\left\{ ||\boldsymbol{x}-\boldsymbol{\mu}||^2 - \sum_{i=1}^{m}((\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\phi}_i)^2 \right\} \tag{9}$$

where

$$\lambda = \frac{1}{n-m}\sum_{i=m+1}^{n} \lambda_i \tag{10}$$

$$= \frac{1}{n-m}\left\{ S - \sum_{i=1}^{m} \lambda_i \right\}. \tag{11}$$

Here S denotes the summation of the diagonal components of Σ. The following function that includes correction term K is proposed for blurred character recognition.

$$\hat{d}_S^2 = \sum_{i=1}^{m} \frac{1}{\lambda_i}((\boldsymbol{x}-\boldsymbol{\mu})^t K\boldsymbol{\phi}_i)^2 + \frac{1}{\lambda}\left\{ ||K(\boldsymbol{x}-\boldsymbol{\mu})||^2 - \sum_{i=1}^{m}((\boldsymbol{x}-\boldsymbol{\mu})^t K\boldsymbol{\phi}_i)^2 \right\}. \tag{12}$$

Eq. 12 is called *modified SMD.*

4 Experiments

We carry on experiments to confirm the effectiveness of the proposed method. As training data, 2,965 kinds of characters used in Section 2.2 are employed. Test data are three sizes of printed characters (6, 7 and 8 point) copied with two modes (thin mode and thick mode) and scanned by an optical scanner. The images are destroyed by copy, especially the ones copied with thick mode. As evaluation functions, SMD and modified SMD are used under the condition of $m = 5$. Experimental results (error rate) are shown in Table 1. In any case, error rate is decreased (or not changed) by modified SMD. The results have shown it is effective to change distribution of image according to degree of blur for blurred Chinese character recognition.

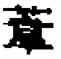
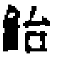
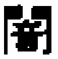
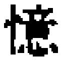
Table 2 shows some examples that are correctly recognized by modified SMD, while they are missed by SMD. In Table 2, original image, correct answer and candidate selected by SMD are displayed. Obviously, each pair of answer and candidate is quite similar, and the blurred part of character is the biggest cause of failure. By using modified SMD, since the information of blurred parts is reduced, comparatively clear parts, no matter they are small or not, will take a more important role in recognition process. For example, in the case of (b), because right

Table 1: Experimental results.

Method	Thin mode			Thick mode		
	6pt	7pt	8pt	6pt	7pt	8pt
SMD	11.4%	2.5%	0.3%	28.6%	15.2%	1.7%
Modified SMD	7.2%	1.5%	0.3%	12.8%	5.5%	1.0%

part of the image is clear, the candidate by SMD is a character that has the same structure in the right. It is correctly recognized by modified SMD because top left part is clear, and this part of answer and candidate is quite different.

Table 2: Examples of images that are correctly recognized by modified SMD.

	(a)	(b)	(c)	(d)
Image				
Answer	葦	飴	闇	憶
Candidate by SMD	繁	胎	簡	撹

5 Conclusions

In this paper, a new recognition algorithm for Chinese characters based on the concept of degree of blur is presented. This method has an especial effect on character images those are not clear. If a character image is blurred, the shape of the character is changed and it always causes the bad influence for recognition. We investigate how the distribution of feature vectors changes if a part of character image is crashed and reconstruct the distribution to respond the change expressed by degree of blur. The proposed method develops the function of the Mahalanobis distance by considering change in distribution. The results of recognition experiments with blurred Chinese character images have shown the effectiveness of the new method.

References

[1] H. Aso, Thinning Algorithm Suitable for Parallel Processing, *IEICE Trans.*, Vol.J76-D-II, No.9, September 1993, pp.2148–2150.

[2] C. J. Hilditch, Linear Skeleton from Square Cupboards, In: *Machine Intelligence 6*, B.Meltzer & D.Michie, Eds., Univ.Press, Edinburgh, 1969, pp.403–420.

[3] E. Oja, Subspace Methods of Pattern Recognition, *Research Studies Press*, 1983.

[4] S. Omachi and H. Aso, Precise Recognition of Blurred Printed Characters, *IEICE Trans.*, Vol.J79-D-II, No.9, September 1996, pp.1534–1542.

[5] T. Saito, H. Yamada, and K. Yamamoto, On the Data Base ETL9 of Handprinted Characters in JIS Chinese Characters and Its Analysis, *IEICE Trans.*, Vol.J68-D, No.4, April 1985, pp.757–764.

[6] F. Sun, S. Omachi and H. Aso, Precise Selection of Candidates for Handwritten Character Recognition Using Feature Regions, *IEICE Trans.*, Vol.E79-D, No.5, May 1996, pp.510–515.

[7] N. Sun, T. Tabara, H. Aso and M. Kimura, Printed Character Recognition Using Directional Element Feature, *IEICE Trans.*, Vol.J74-D-II, No.3, March 1991, pp.330–339.