CASE 79

Character Recognition Using the Mahalanobis Distance

Abstract: Today, the demand for electronic document processing is becoming stronger and stronger. However, technologies to recognize handwriting have not been well developed. In this study, improvement in handwriting recognition, to recognize a pattern or multidimensional information, was studied using the Mahalanobis–Taguchi system (MTS). It has been confirmed that the use of MTS is effective for multidimensional information.

1. Introduction

In order for a computer to recognize a handwriting character by taking advantage of a multidimensional information process, we create a base space. Considering that the base space needs to recognize various types of commonly used characters, we collected as many writings for a particular character but different shapes as possible by asking a number of testees to hand-write characters. Each testee writes down a *hiragana* character, or Japanese syllabary, on a 50 mm by 50 mm piece of paper as a sample character. The total number of writings amounted to approximately 3000. Figure 1 shows a sample character.

A sample character is saved as graphical data according to the procedure shown in Figure 2, and its features are extracted. The graphical data are divided into $n \times n$ cells (Figure 3). If a part of the character lies on a certain cell, the cell is set to 1. Otherwise, the cell is set to 0. This process is regarded as binarization of the graphical data. Since the number of binarized cells is $n \times n$ without special treatment, if n is equal to 50, the number adds up to 2500. In such a case it is quite difficult to calculate a 2500 × 2500 inverse matrix with a personal computer, which is required to create a base space. Therefore, by extracting the character's features, we reduce the number of binarized cells.

For the 25 × 25 binary data for the character shown in Figure 4, we performed feature extraction in accordance with the following three procedures:

1. *Integral method I.* In each row and column, we summed up the number of consecutive cells, including empty cells, from the cell with a character data item (or 1) to the one with a character data item.
2. *Integral method II.* In each row and column, we simply counted the number of cells with a character data item.
3. *Differential method.* In each row and column, we counted the number of groups of a single cell or consecutive cells with a character data item.

The rightmost columns and bottom rows in Figure 4 show the values calculated by the three types of feature extraction methods above.

2. Procedure for Character Recognition

Figure 5 summarizes the procedure, starting from the collection of a handwriting character through the extraction of features and creation of a base space, to recognition of a character.

Without any special modification, handwriting characters were gathered from as many people as possible. The characters collected were read with a scanner and saved as graphical data. These graphical data were divided into $n \times n$ cells, converted into 0/1 data, and their features were extracted.

Taguchi's Quality Engineering Handbook. Genichi Taguchi, Subir Chowdhury and Yuin Wu

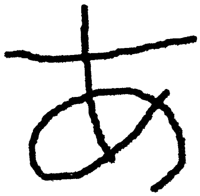

Figure 1
Handwriting the character "ah"

Setting the features extracted to our basic data, we determined a base space and computed the Mahalanobis distance for each of the characters collected. As a nest step, we gathered new characters irrelevant to creation of the base space. For these new characters, we calculated the Mahalanobis distances and performed character recognition for their target characters.

3. Preparation of Base Space for Handwriting Character

Here we focused only on the case of using the character "ah" as a base space. Preparing four character groups: the character "ah" used as the base space, another "ah" (not used for the standard space), and "oh" and "nu," which look similar to "ah." We calculated the Mahalanobis distances for the standard space created by the character "ah" from the first group. Figure 6 shows the distribution of the Mahalanobis distance for each character. Using Figure

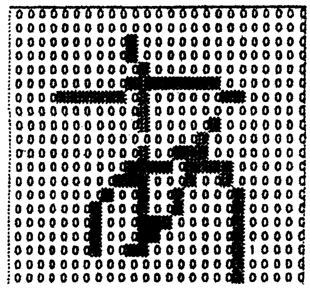

Figure 3
Binarization of handwriting character in Figure 1

6 we investigated the probability of recognizing an unknown character. The horizontal and vertical axes indicate a Mahalanobis distance and frequency, respectively. The Mahalanobis distances for the character used for the base space are distributed on both sides of 1, which is their average. The distances for the second group of the character "ah" are almost fully contained in the base space. On the other hand, although both of the distributions of the distances for "oh" and "nu" to some extent overlap the base space around 1.5, we are likely to discriminate them from the character "ah" as different characters. Ideally, the second group of "ah," which was not used to construct the base space, should overlap the first group of "ah," which was used to construct the base space.

Although the number of "ah" characters used to create a base space had initially been set to 1500, by increasing the number to 3000 we sharpened the distribution and made the discrimination easier. On

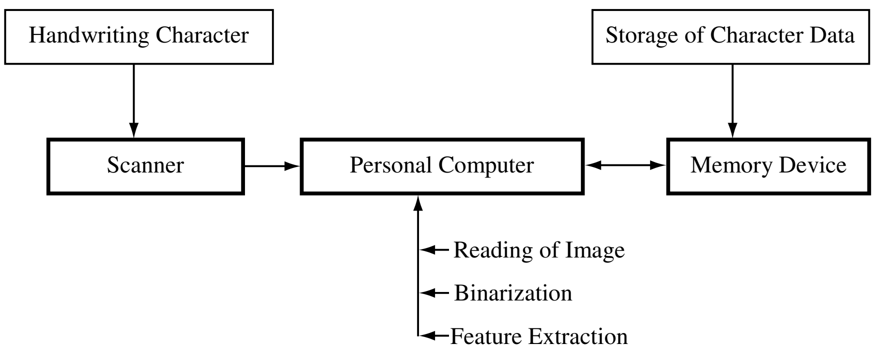

Figure 2
Reading of handwriting characters and extraction of features

																										Integral Method I	Integral Method II	Differential Method
	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	9	9	1
	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	25	19	2
	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	11	2	2
	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	1	1	0	0	0	0	11	6	3
	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	10	8	3
	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	0	11	7	2
	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	1	1	1	1	0	0	0	0	15	8	3
	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	17	6	4
	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	0	0	19	7	4
	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	20	5	4
	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	0	22	6	4
	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	22	6	4
	0	1	1	1	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	1	1	0	23	9	4
	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0	0	22	5	3
	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	22	6	3
	0	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	21	7	3
	0	0	1	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	0	0	0	19	10	3
	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	14	8	2
	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1
Integral Method I	1	18	19	19	19	20	20	21	21	21	21	23	20	18	17	16	15	21	20	19	19	18	17	15	1			
Integral Method II	1	5	7	7	4	5	4	4	5	4	11	15	8	4	5	4	7	7	6	5	7	5	7	4	1			
Differential Method	1	2	3	3	3	3	3	3	3	3	2	3	2	3	3	3	3	3	3	4	4	3	3	2	1			

Figure 4
Japanese character whose features are extracted in 25 × 25 cells

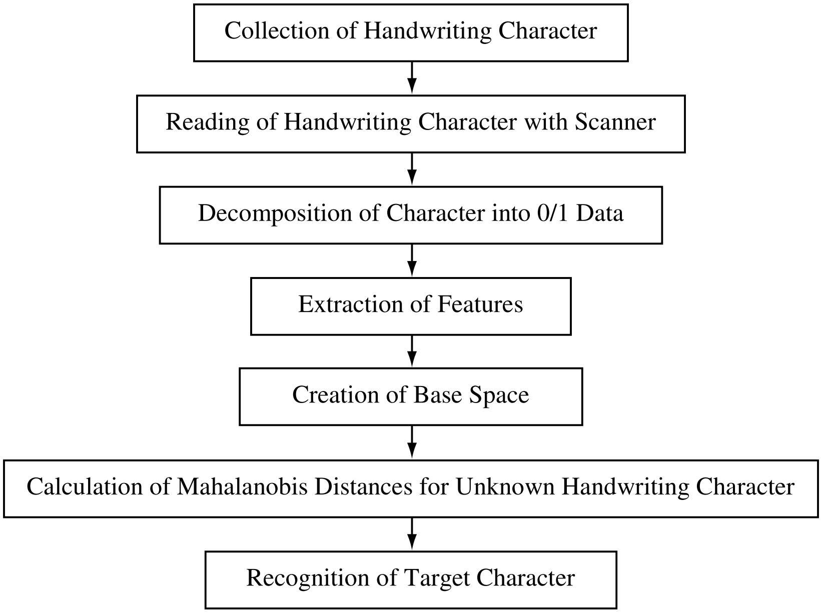

Figure 5
Procedure for handwriting character recognition

the other hand, the total number of "oh," "nu," and "ah" that were not used for the base space is 100.

To clarify the recognition rate of target characters according to Figure 6, for the character "ah" that is not used for the base space, we plotted the relationship (1 − cumulative frequency) versus Mahalanobis distance, whereas for "oh" and "nu" the relationship of cumulative frequency versus Mahalanobis distance is drawn. Both of the results are shown in Figure 7.

If the Mahalanobis distance of a target character is less than 1, we can judge it as a character "ah" without fail. However, when it becomes more than 1, we cannot reject the probability that the target character should be "oh" or "nu," which look similar to "ah." On the other hand, if the Mahalanobis distance of the target character exceeds 3, the likelihood that the character is "ah" diminishes to almost zero. For characters that do not resemble "ah" in the slightest, we do not experience difficulty with recognition.

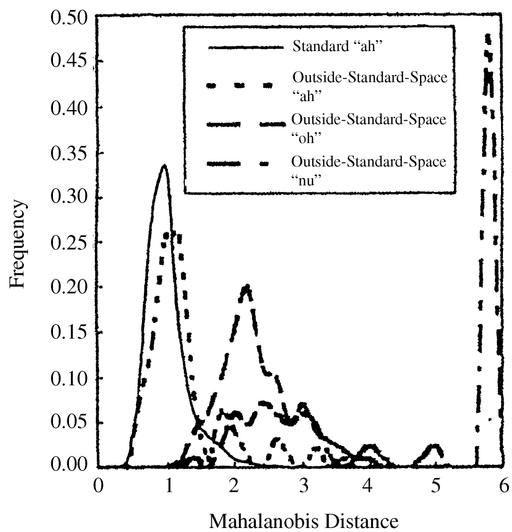

Figure 6
Distribution of characters by the Mahalanobis distance

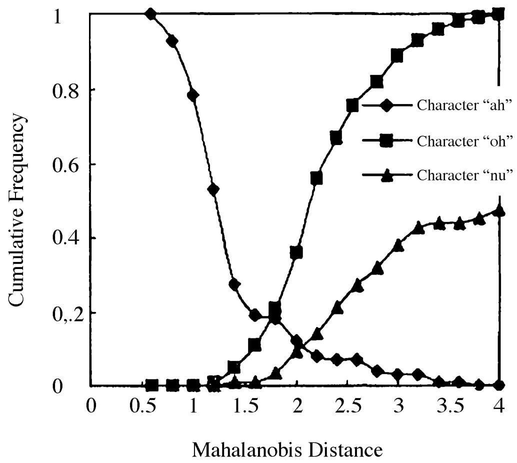

Figure 7
Relationship between Mahalanobis distance and cumulative frequency using base space for "ah"

Table 1
Frequency of occurrence of target character for base space of character "ah"

(*a*) Occurrences of "ah" and "oh"

	Mahalanobis Distance	
Character	Less Than 1.5	More Than or Equal to 1.5
"ah"	0.8	0.2
"oh"	0.1	0.9

(*b*) Occurrences of Combinations of "ah" and "oh"

	"ah"	
"oh"	Less Than 1.5	More Than or equal to 1.5
Less than 1.5	0.08	0.02
More than or Equal to 1.5	0.72	0.18

4. Conclusions

Table 1 shows the frequencies of occurrence for "ah," "oh," and "nu" when the Mahalanobis distance = 1.5. Using (*a*) of Table 1, we obtain (*b*) for the character recognition rate, which indicates 0.72 as the probability of correct recognition and 0.02 as that of incorrect recognition. The hatched areas represent the probabilities that the Mahalanobis distances for "ah" and "oh" above and below 1.5. Comparing the Mahalanobis distances calculated for "ah" and "oh," we make a correct judgment if the distance for "ah" is smaller than that for "oh," an incorrect judgment if for "ah" is larger than that for "oh." Since the probability of correct recognition in the hatched areas can be assumed to be more than half the total probability, or 0.13, by adding up this correct recognition rate and the aforementioned 0.72, we obtain over 90% as the overall recognition rate.

Reference

Takashi Kamoshita, Ken-ichi Okumura, Kazuhito Takahashi, Masao Masumura, and Hiroshi Yano, 1998. Character recognition using Mahalanobis distance. *Quality Engineering*, Vol. 6, No. 4, pp. 39–45.

This case study is contributed by Takashi Kamoshita, Ken-ichi Okumura, Kazuhito Takahashi, Masao Masumura, and Hiroshi Yano.