Printed Character Recognition by Matching Partial Patterns

Akira KUREMATSU* and Seiichi INOUE**

Abstract

A method to recognize printed characters by matching partial patterns is introduced. Those partial patterns like line segments, various corners, or separated segments are extracted from sample patterns and stored in a memory. Values of similarity are calculated by correlating those partial patterns and small regions of input patterns at predetermined positions. The scores are counted on the basis of similarity value and weight value for each character. The input pattern is determined to be the character which gives the maximum sum of scores. The advantages of this method are that it is not strongly affected by the noises like spots and grains and the required memory capacity is smaller than that needed for the whole pattern matching method. Results of simulation for 40 alphanumeric and symbol characters of ISO-B font are shown.

1. Introduction

Some optical character readers for printed characters have already been used practically. However, further study is needed to develop more economical readers having fewer limitations as for the quality of input printed characters. The presently developed printed character recognition methods may be broadly classified into the stroke analysis method [1][2], the partial figure identifying method [3], and the whole character pattern matching method [4]. In the stroke analysis method, basic line elements which compose a pattern are prepared, and the pattern is identified by the positions where the basic line elements exist. This method is suitable for a character mainly consisting of lines, and hence, the number of identifiable character of type is limited. In the partial figure identifying method, a character is divided into two-dimensional grids; partial figures such as strokes, angles, spaces between strokes are observed in the figure of the character; the character is recognized

This paper first appeared in Japanese in Joho-Shori (Journal of the Information Processing Society of Japan), Vol. 17, No. 12 (1976), pp. 1113~1119.

* Research and Development Laboratories, Kokusai Denshin Denwa Co., Ltd.
** Kokusai Denshin Denwa Co., Ltd.

through the combinations of the partial figures. The logics to recognize a character correctly regardless of noises such as a deviated position of the character and contamination of the paper or character are complicated. In the pattern matching method, the position of the characters and the thicknesses of the lines must be rigidly standardized, and an increase in the memory capacity cannot be avoided because from one to about five standard figures must be stored into the memory for each character. This paper introduces a new printed character recognition method which combines the merits of the whole character pattern matching method and the partial figure recognition method [5][6][7].

2. Method of Recognition by Matching Partial Patterns

Those partial figures such as spaces between strokes, angles, and strokes which compose a character and which are useful in recognizing that character are called partial feature patterns. A partial feature pattern is expressed in three values: a black portion, a white portion and an indefinite portion which may be either black or white. An input pattern is placed over a partial feature pattern at a position predetermined for each partial feature pattern. The same partial feature pattern may be used at two or more positions. Fig. 1 shows examples of partial feature patterns.

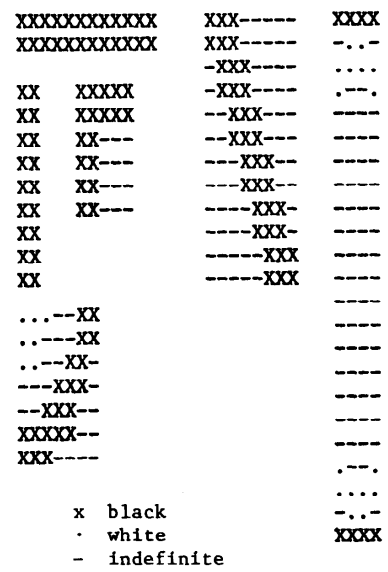

Fig. 1 Examples of partial feature patterns

The degree of matching of a partial feature pattern and a character pattern is calculated for similarity. Similarity is defined as follows, with the similarity of a character pattern against a partial feature pattern k expressed as S_k ;

$$S_k = \frac{\sum_{x}^{D_k}\sum_{y} F_k(x,y)\ P(x,y)}{\sum_{x}^{D_k}\sum_{y} \left| F_k(x,y) \right|} \qquad (2.1)$$

where, D_k : Domain of partial feature pattern

$F_k(x,y)$: Function which expresses the partial feature pattern

(When the grids (x,y) are in black level, $F_k(x,y) = 1$;

when the grids (x,y) are in white level, $F_k(x,y) = -1$; and

when the grids (x,y) are indefinite, $F_k(x,y) = 0$.)

P(x,y): Function which expresses the character pattern

(When the grids (x,y) are in black level, $P(x,y) = 1$; and

when the grids (x,y) are in white level, $P(x,y) = -1$.)

Input characters are separated after being photoelectrically transformed, and entered into a two-dimensional register after being quantized into two values. The two-dimensional register centers the character pattern. To prevent reduction of similarity due to position deviation, the character pattern on the register is shifted vertically and horizontally within the permissible range, and the maximum value of similarity of the character pattern and the partial feature pattern obtained through the shifting is used as the feature value of the partial feature pattern of the input character. The feature value of each partial feature pattern of the input character is compared to the feature weight value of each partial feature pattern predetermined for each character, providing each partial feature pattern with a score. Thus, the character having the largest sum score is selected.

First of all, the feature value (f_k) of each partial feature pattern (k) is calculated for an input pattern. Next, scores are obtained using with the function $h(f_k, \omega_{ik})$ which is determined by the relationship between the feature weight value (ω_{ik}) of each character (i) for the partial feature pattern (k) and the feature value (f_k). For the feature weight value, the similarity of the partial feature pattern to the standard pattern is used, because it is to be suitable for the pattern matching. An h is called a score function for identification. Fig. 2 shows an example. An h is used as a ternary function instead of a binary function, since a better recognition rate is obtained for a ternary function. The sum of scores for each character is calculated as shown below, and the character having the largest G_i is selected.

$$G_i = \sum_{k=1}^{N_F} h(f_k, \omega_{ik}), \quad i = 1 \sim N_c \tag{2.2}$$

where, N_F : Number of partial feature patterns

N_c : Number of types of characters

N_F is used as the required number of partial feature patterns, with the recognition rate attained above a certain value.

When recognizing a printed character by matching a partial pattern, it is important to select the most appropriate types and shapes for partial feature patterns. A computer aided procedure that an operator properly aids and intervenes in setting up partial feature patterns was used. Details are described in [7].

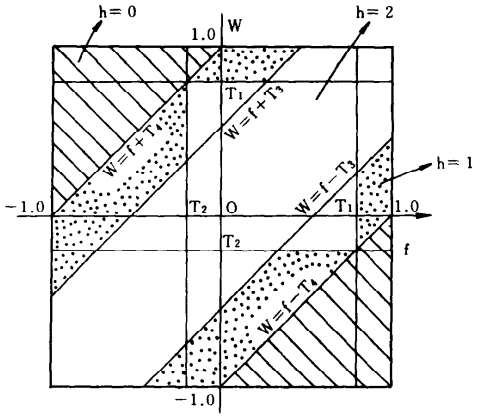

Fig. 2 Score function in the recognition where f is the feature value and w is the weight of the feature

3. Simulation

Simulation of recognition was conducted for 40 typewritten characters (36 alphanumeric characters and four symbols (. , / -) of ISO font B). A standard pattern was prepared by piling three types of character patterns having slight differences in thicknesses of the printed characters. The character patterns were divided into 28 grids vertically and 20 grids horizontally per character. The total number of partial feature patterns is 42 and memory capacity occupied by the partial feature patterns is 2181 bits, which is, about 1/10 that needed to memorize a similar number of whole character patterns.

Typewriting pressure was reduced gradually to reduce the thickness of the prints gradually, and thus, ten sets of characters (total 400 characters) in ten different steps of thickness were prepared. Table 1 shows the results of recognitions for the ten sets of characters. The larger sample numbers indicate decreasing thickness. Those having one or fewer score

Table 1 Results of recognition of sample patterns

SAMPLE NUMBER	1,2	3,4	5,6	7,8	9,10	TOTAL
ERROR	0	0	0	0	0	0
REJECTION	0	0	3 {0,0,5}	0	0	3

Table 2 Results of recognition where standard patterns are shifted

POSITION	RESULT OF RECOGNITION	
	ERROR	REJECTION
1	0	2 {B,D}
2	0	0
3	0	1 {4}
4	0	0
5	0	0
6	0	1 {0}
7	0	1 {0}
8	0	4 {H,O,0,8}

NUMBER OF SHIFT POSITION

8	5	3
7		2
6	4	1

differences were rejected. Those patterns shown in Table 1 as rejected patterns were recognized correctly but the score differences were one.

Further, to obtain an estimate of the recognition ability of the recognition system by matching partial patterns, the patterns were shifted vertically and horizontally one grid against the standard pattern. Table 2 shows the result of recognition.

4. Conclusion

A new method of printed character recognition by matching partial feature patterns has been described. As far as those 40 alphanumeric and symbol characters of ISO font B are concerned, it is possible to recognize them with a recognition ability only slightly lower than that of the whole character matching method. There exist conditions requiring preprocessing. The permissible range of position deviation should be within one grid of the standard pattern and influence of deformation due to random noise is small. For characters which are difficult to recognize by this method, it is believed that a recognition ability equivalent to that of the whole character matching method can be achieved by setting the whole pattern as the partial pattern because the number of types of such characters is small.

Acknowledgement

The authors express their deep appreciation to Dr. Y. Nakagome and Dr. H. Kaji for their support and encouragement during this study.

References

[1] R. L. Grimsdale et al., "A System for the Automatic Recognition of Patterns", Proc. of IEE, Vol. 106, No. 26, pp. 210-221, (1959).

[2] G. L. Fischer, "Optical Character Recognition", Spartan Books, (1962).

[3] R. B. Hennis, "The IBM 1975 Optical Page Reader", IBM Journal Res. Dev., Vol. 12, No. 5, pp. 346-371, (1968).

[4] Iijima, Mori, "OCR attaining human ability -- ASPET/71", NIKKEI ELECTRONICS, No. 30, pp. 66-80, (1972).

[5] A. Kurematsu, S. Inoue, "Recognition of Printed Character by Matching Partial Pattern", Technical Report of the Research Group on Pattern Recognition and Learning of the Institute of Electronics and Communication Engineers of Japan, PRL 73-91, (1974).

[6] A. Kurematsu, S. Inoue, "Simulation of Printed Character Recognition by Matching Partial Patterns", Technical Report of the Institute of Electronics and Communication Engineers of Japan, PRL75-13, (1975).

[7] A. Kurematsu, S. Inoue, "Printed Character Recognition by Matching Partial Pattern", Jour. of Information Processing Society of Japan, Vol. 17, No. 12, (1976).