A New Approach for Recognition Multifont Chinese Characters Used in a Special Application

Ling–Hwei Chen and Yeuan–Kuen Lee

Institute of Computer and Information Science
National Chiao Tung University, Taiwan, Republic of China

Abstract

In this paper, a new approach for recognizing a limited set of printed characters is provided. It can be used in some special applications, which require a real time response and high recognition rate.

1: Introduction

Chinese character recognition is a very hard problem because of the following difficulties : a large character set; complicated structure for individual character; and many groups of similar characters. Many approaches [1–2] have been developed for recognizing all commonly used characters. They are usually complex. Some applications do not need to recognize all of the Chinese characters. For examples, Chinese cheque, receipt, etc. Based on this reason, here we propose a simpler algorithm, which satisfies real time process requirement and high recognition rate, to recognize a limited set of multifont Chinese characters used in a special application.

Stroke is one of the major features of Chinese characters. Since the aim of the proposed method is to recognize a special set of printed Chinese characters, only the horizontal/vertical strokes and the crossings among these two types of strokes are used as the features of characters. In the learning stage, a random model for each type of characters is established. In the recognition stage, the character model which has the minimal error with the input pattern is considered as the recognition result. The proposed recognizer is insensitive to noise. Some experimental results show that the method exactly provides high recognition rate and has high computing speed.

2: Feature extraction

As mentioned above, the horizontal and vertical strokes are used as the features of printed characters. To avoid many small horizontal (resp. vertical) segments existing in the same horizontal (resp. vertical) line being considered to be a horizontal (resp. vertical) stroke, we provide a modified Hough transform to find the horizontal and vertical strokes, i.e., two specific kinds of lines $y = p$ and $x = p$. Hence, only two one–dimensional accumulators $AH(y)$ and $AV(x)$ for horizontal and vertical strokes, respectively, are needed in the parameter space. For each point (x, y) in one character pattern CP, two operations are conducted:

$$AH(y) = AH(y) + 1, \text{ if } \begin{matrix}(x-2,\ y) \in CP \\ \text{and } (x+2,\ y) \in CP;\end{matrix}$$

$$AV(x) = AV(x) + 1, \text{ if } \begin{matrix}(x,\ y-2) \in CP \\ \text{and } (x,\ y+2) \in CP.\end{matrix}$$

After applying the modified method, the strokes can be extracted by finding the local maxima of accumulators $AH(y)$ and $AV(x)$. Then, each character pattern is normalized into 60x60, and the normalization operation is only applied to those extracted strokes. After extracting all horizontal and vertical strokes, we can record the crossings among these strokes.

0-8186-2915-0/92 $3.00 © 1992 IEEE

3: Stroke matching and learning

In this section, we will present the methods used in the learning stage. First, an error measure E_1 for two strokes of the same type is defined, one in a new learning pattern, the other in the reference model. There are three cases for stroke matching error.

Case 1: For a stroke S in the learning pattern, if there does not exist one stroke in the reference model to match S, set $E_1(S) = 0.75$.

Case 2: For a stroke S in the reference model, if there does not exist one stroke in the learning pattern to match S, set E_1 to be its associated probability.

Case 3: For a stroke S in the reference model, there exists one in the learning pattern to match S, set

$$E_1(S)=2*(1-SS*\overline{CS}).$$

Note that SS is the stroke probability and $\overline{CS}$ is the mean crossing similarity.

The mean crossing similarity is the mean of the similarity values of all crossings in two matched strokes. Let PST be the set of strokes in the learning pattern, MST be the set of strokes in the reference model, and let

$$PMS = \{ PS \mid PS \in PST, \text{ and } \exists \text{ one } MS \in MST \text{ matches } PS \}.$$

Let $PS \in PST$, $MS \in MST$. Assume that PS matches MS. Let PC be the set of all crossings in PS, MC be the set of all crossings in MS. Let C1 = (PS, PS') and C2 = (MS, MS') be two crossings belonging to PC and MC, respectively. We say that C1 matches C2, if PS' matches MS'. Note that $C = (S_1, S_2)$ means that C is the crossing of strokes S_1 and S_2.

Let CS(C) be the similarity measure of the crossing C. There are five cases for the similarity measure of one crossing.

Case 1: For C1, if PS' matches MS', set CS(C1) = probability of C2 ;

Case 2: for C1, if PS' matches a stroke MS1, $MS1 \in MST$, and there does not exist a crossing (MS, MS1) in MC, set CS(C1)=0;

Case 3: for C1, if PS' does not match any stroke of MST, set CS(C1) = 0;

Case 4: for C2, if MS' matches a stroke PS1, $PS1 \in PST$, and there does not exist a crossing (PS, PS1) in PC, set CS(C2) = 1 – probability of C2.

Case 5: for C2, if MS' does not match any one of PST, we don't count the crossing simularity for C2. All of the crossings in the case are grouped into a set named MCD. Let

$$MPC = \{ C2 \mid C2 \in MC, \text{ and } \exists\ C1 \in PC, \text{ s.t. } C1 \text{ matches } C2 \},$$

and let n_{MPC} be the number of crossings in MPC and n_{MCD} be the number of crossings in MCD. Based on the above-introduced measure, we define $\overline{CS}$ as

$$\overline{CS}=\Big(\sum_{C1\in PC} CS(C1) + \sum_{C2\in MC-MPC-MCD} CS(C2)\Big) / (n - n_{MPC} - n_{MCD}),$$

where n is the total number of PC and MC.

And the total matching error ($\overline{E_1}$) between a learning pattern and its reference model can be defined as

$$\overline{E_1} = \sum_{PS\in PST} E_1(PS) + \sum_{MS\in MST-PMS} E_1(MS).$$

Based on this measure, an optimal stroke matching method, under the minimum error sense, for a new learning pattern and its reference model is then provided. The method does an exhaustive search. By applying the matching algorithm to all the learning patterns of the same character type, an individual random reference model is established for each type of characters. Note that a random reference model contains the information of strokes and crossings. Each stroke (or crossing) has a probability value associated with it. Assume that a random model is established by n learning patterns, and a stroke **S** exists in k learning patterns, then the associated probability value of **S** in the random model is k/n. The definition for that of a crossing is similar.

4: Recognition

In the recognition stage, a new error measure is proposed. Let CP be an input character pattern, CM be the reference model. Let PH and PV be the sets of horizontal and vertical strokes in CP, MH and MV be the sets of horizontal and vertical strokes in CM, respectively. Let

$$PH = \{ p_0, p_1, p_2, \cdots p_l \};$$
$$PV = \{ q_0, q_1, q_2, \cdots q_m \};$$
$$MH = \{ r_0, r_1, r_2, \cdots r_n \};$$
$$MV = \{ s_0, s_1, s_2, \cdots s_o \}.$$

And let E_r be the matching error between the stroke of CP and the stroke of CM. Consider the following cases:

Case 1: p_i can not match any one of MH, $0 \le i \le l$, or q_j can not match any one of MV, $0 \le j \le m$. Set $E_r(p_i) = 3$, or $E_r(q_j) = 3$.

Case 2: r_i can not match any one of PH, $0 \le i \le n$, or s_j can not match any one of PV, $0 \le j \le o$. Set $E_r(r_i) = Pr(r_i)$ or $E_r(s_j) = Pr(s_j)$.

Case 3: p_i matches r_j, for some j, or q_i matches s_j, for some j. Set

$$E_r(p_i) \text{ (or } E_r(q_i)) = \overline{CE}(p_i, r_j)$$
$$\text{(or } \overline{CE}(q_i, r_j)).$$

Where $\overline{CE}(p_i, r_j)$ is the total crossing error in the strokes p_i and r_j. Let (h, v) be the crossing of horizontal stroke h and vertical stroke v, and Pc(h, v) be the probability of crossing point (h, v). Assume that a horizontal stroke p_i matches r_j. The crossing error CE for a crossing in a stroke can be defined as follows :

Case A: (p_i, q_l) exists, q_l matches s_k, and (r_j, s_k) exists. Set $CE(p_i, q_l) = 0$ and $CE(r_j, s_k) = 0$;

Case B: (p_i, q_l) exists, q_l matches s_k, but (r_j, s_k) doesn't exist. Set $CE(p_i, q_j) = 1$;

Case C: (p_i, q_l) exists, but q_l doesn't match any stroke in MV. Set $CE(p_i, q_l)=0$;

Case D: (r_j, s_k) exists and q_l matches s_k, but (p_i, q_l) doesn't exist. Set $CE(r_j, s_k) = Pc(r_j, s_k)$.

Case E: (r_j, s_k) exists, but s_k doesn't match any stroke in PV. Set $CE(r_j, s_k) = 0$.

Note that for the vertical strokes, the crossing error is similarly defined. Based on these definitions, we can define the total

crossing error $\overline{CE}(p_i, r_j)$ for strokes p_i and r_j as

$$\overline{CE}(p_i, r_j) = \Sigma\, CE(p_i, q_l) + \Sigma\, CE(r_j, s_k),$$

where (p_i, q_l) is one of the crossings in stroke p_i, (r_j, s_k) is one of the crossings in stroke r_j.

Let

$$MPH = \{\, r_i \in MH \mid \exists\ a\ p_i \in PH, \text{ s.t. } p_i \text{ matches } r_i \,\},$$

$$MPV = \{\, s_i \in MV \mid \exists\ a\ q_i \in PV, \text{ s.t. } q_i \text{ matches } s_i \,\}.$$

Then we can evaluate the matching error $\overline{E_r}$ between pattern CP and model CM as

$$\overline{E_r}(CP, CM) = \sum_{p_i \in PH} E_r(p_i) + \sum_{q_i \in PV} E_r(q_i) + \sum_{r_i \in MH-MPH} E_r(r_i) + \sum_{s_i \in MV-MPV} E_r(s_i).$$

After evaluating all the matching errors of pattern CP with all models CM, we then select one class model having the minimal error as the recognition result.

In the learning stage, an optimal matching algorithm is used to find the best matching. In the recognition stage, we use a nearly–optimal matching algorithm [3] to save the searching time.

5: Experimental results

In order to demonstrate the feasibility of the proposed algorithm, we test our proposed approach on the set of 17 characters which appear on the Chinese cheque. In the learning stage, there are 7 different fonts and only one sample of each font is trained. 21 printed Chinese characters for each class are used to test and all of them have been recognized correctly. The experimental results show that the recognition rate is high.

References

[1] P. N. Chen, Y. S. Chen and W. H. Hsu, " Stroke relation coding — A new approach to the recognition of multi–font printed Chinese characters," Int. J. of Pattern Recognition and Artificial Intelligence, Vol. 2, No. 1, 1988, 149–160.

[2] S. Zhang, B. Taconet and A. Faure, " A complexity measure based algorithm for multifont Chinese character recognition," Proc. 10th Int. Conf. on Pattern Recognition, 1990, 573–577.

[3] Y. K. Lee, "A new approach for a small set of multi–font Chinese Character Recognition," Master Thesis, Institute of Computer and Information Science National Chiao Tung University, Taiwan, Republic of China, June 1991.