Printing Press Fuzzy Identification based on Pixel Distribution Probability of Character Image

Ning Wang
School of Medical Information Engineering,
Guangdong Pharmaceutical University, Guangzhou, Guangdong, 510006, China
gz0609@sina.com

Abstract

Printing press identification can be realized with comparison between the same printed characters by different printing presses. This paper proposed a new method of printing press identification. It takes advantage of the method of Casey and Nagy and overcomes the disadvantage of the traditional the method of character recognition. In this method, the pixel distribution probability of character image is the basal index. The minimal distance product and the maximal fuzzy correlation measure are the double-optimized evaluation indexes. The precision of this method is on the pixel level. The tiny difference of same printed characters by different printing presses can be identified by the difference of pixel of character image. So the pixel distribution probability of character image is different too. With the experiment of printing press identification, the accurate identification rate is about 96.18%. The result of experiment showed printing press identification based on pixel distribution probability is effective and feasible.

1. Introduction

With the development of printing technology, the presswork is used widely. In some cases and analysis of quality of printing press, the type of printing press needs to be identified. But it is difficult to observe and measure the difference between printed same characters by different printing presses. The printed condition, system error and other factors will bring some problems to get accurate data. The same printed characters by different printing presses need to be compared to identify the type of printing presses. With classical method, the structural feature of character must be extracted [1]. But the feature of same characters is same. Therefore, the classical method of character recognition is unfit for printed character identification [2][3].

For eliminating these interferential factors effectively, the concept of similarity of printed character image can make the problem of character identification become simple. Now, we only need to compare the similar measure between images to evaluate the most similar printing press. In mathematics, correlation coefficient often is used as the model of similarity. Otherwise, the evaluation of similar measure can be done by fuzzy comparison. So fuzzy synthetic evaluation based on maximum membership principle and correlation coefficient is a suitable mathematic model for printing press identification. It is not complicated to be processed by computer.

This research studied a new method of character identification according to the domestic and overseas character recognition technology. It uses the distance and similarity of indexes, and takes advantage of the method of Casey and Nagy which computed pixel directly to match template. The key of this method is that the pixel distribution probability of character image is used as the basal index of character identification. The minimal distance product and the maximal fuzzy correlation measure of pixel distribution probability are the double-optimized evaluation indexes to identify the different printing press. The precision of this method is on the pixel level. The tiny difference of same printed characters by different printing presses can be reflected by the difference of pixel of character image. The pixel distribution probability is different too. The pixel distribution probability of character image is a holographic index because it includes the all structural and statistic information of character [4]. With above three indexes, printing press identification can be finished well.

2. Printed character image acquiring and processing

All printed characters are transformed into the format of bitmap by professional scanner in same parameter. These bitmaps are binarized by improved form-analysis of histogram [5][6].

978-0-7695-3490-9/08 $25.00 © 2008 IEEE
DOI 10.1109/PACIIA.2008.356

3. Character image segmentation and index setup

These Bitmaps must be segmented on several modes before these characters are identified. For example, they can be segmented into 16 equal-area quadrate sections on three modes—gridding, horizontal bar and vertical bar, as illustrated in Figure 1, 2, and 3 [7].

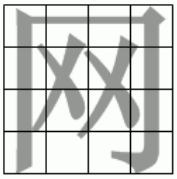

Figure 1 The 16 equal-area quadrate sections on 4×4 gridding segmentation mode

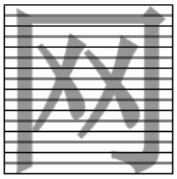

Figure 2 The 16 equal-area quadrate sections on 16×1 horizontal bar segmentation mode

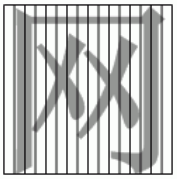

Figure 3 The 16 equal-area quadrate sections on 1×16 vertical bar segmentation mode

Let p_{ij} be the pixel distribution probability of strokes of every section. It is the ratio of the area of strokes of every section to the area of character strokes. The area of strokes is expressed by the pixels of strokes. We have the following formula to calculate p_{ij}.

$$p_{ij} = n_{ij} / n \quad (i = 1,2,\cdots,m \quad j = 1,2,\cdots,n) \tag{1}$$

$$^{*}p_{ij} = {}^{*}n_{ij} / {}^{*}n \quad (i = 1,2,\cdots,m \quad j = 1,2,\cdots,n) \tag{2}$$

Where n is the total pixels of character, n_{ij} is the pixels of strokes of section, i expresses the different segmentation mode, j expresses the sequence and number of section.

The letter with superscript "*" expresses the identified character.

4. The minimal distance product of pixel distribution probability of character image

The absolute distance of pixel distribution probability between standard character and identified character is a simple and effective index. It reflects the difference of whole character. The product of absolute distance expresses the conjunct and intersectant effect of multi-segmentation. It increases the difference of filtered standard characters and makes the character evaluation easier. It is also the standard of the primary character identification.

The minimal distance product of pixel distribution probability can be calculated by formula (3).

$$Q = \prod_{j=1}^{n}\left(\sum_{i=1}^{m}\left|p_{ij} - {}^{*}p_{ij}\right|\right) \quad (i = 1,2,\cdots,m \quad j = 1,2,\cdots,n) \tag{3}$$

$$Q_{\min} = \{Q_k\} \qquad (k = 1,2,\cdots,g) \tag{4}$$

The minimal distance product is key index for printing press identification. Sometimes, the primary identification based on the minimal distance product can get the accurate result of character identification. 6 characters that Q is minimal are selected for printing press fuzzy identification.

5. The maximal fuzzy similarity measure of pixel distribution probability

After primary identification, 6 characters that Q is minimal are selected for fuzzy identification. Correlation coefficient reflects the discrete degree of data and quantificationally gives the information of accuracy. Therefore, the fuzzy synthetic evaluation based on correlation coefficient of pixel distribution probability is the ultimate printing press identification [8].

5.1. The correlation coefficient of pixel distribution probability and fuzzy membership

The template matching is the basic principle of this method. When two printed characters completely match, their correlation coefficient of pixel distribution probability should be 1. So the two characters will be thought to be printed by the congeneric printing press. In fact, with the error and disturbance, correlation coefficient is always between 0 and 1. It is the index of fuzzy membership and can express the similar degree of two characters. So, it also reflects the fuzzy relationship between two set of characters. The correlation coefficient of relevant pixel distribution probability can be calculated by formula (5) [9].

$$r_i = \frac{\left|\sum_{j=1}^{q}\left(p_{ij}-\overline{p_i}\right)\left(*\,p_{ij}-*\overline{p_i}\right)\right|}{\sqrt{\sum_{j=1}^{q}\left(p_{ij}-\overline{p_i}\right)^2}\sqrt{\sum_{j=1}^{q}\left(*\,p_{ij}-*\overline{p_i}\right)^2}} \quad (i=1,2,\cdots,m \quad j=1,2,\cdots,q) \tag{5}$$

5.2. The distribution of correlation coefficient

According to formula (5), the correlation coefficient of all printed characters is calculated on different modes. Table 1 shows the distribution of calculated correlation coefficient in compared printed characters B_i.

Table 1 The distribution of correlation coefficient

	B_1	B_2	⋯	B_n
A_1	r_{11}	r_{12}	⋯	r_{1n}
A_2	r_{21}	r_{22}	⋯	r_{2n}
⋯	⋯	⋯	⋯	⋯
A_m	r_{m1}	r_{m2}	⋯	r_{mn}

5.3. Fuzzy synthetic evaluation

Let $\mu_A(a_i)$ be the membership function of index, $\mu_B(b_h)$ be the membership function of evaluation, $\boldsymbol{B}$ is evaluation matrix, $R(a_i, b_h)$ be the membership function of fuzzy relationship between $\mu_B(b_h)$ and $\mu_A(a_i)$, and $\boldsymbol{R}$ be the fuzzy relationship matrix. So printing press fuzzy synthetic identification can be finished by the following formula [10].

$$\mu_B(b_h) = \underset{\max}{SUP}\left(\mu_A(a_i) \wedge R(a_i, b_h)\right) \quad (i=1,2,\cdots,m \quad h=1,2,\cdots,n) \tag{6}$$

The weight of index is set according to segmentation mode. The weight coefficient needs to be normalized. It is the mean value of correlation coefficient of single index.

$$A_1 = \overline{r_1}\Big/\sum_{k=1}^{m}\overline{r_k}\ ,\ A_2 = \overline{r_2}\Big/\sum_{k=1}^{m}\overline{r_k}\ ,\ \cdots,\ A_m = \overline{r_m}\Big/\sum_{k=1}^{m}\overline{r_k}\quad (k=1,2,\cdots,m) \tag{7}$$

Where

$$\overline{r_1} = \frac{1}{n}\sum_{h=1}^{n} r_{1h}\ ,\ \overline{r_2} = \frac{1}{n}\sum_{h=1}^{n} r_{2h}\ ,\ \cdots,\ \overline{r_n} = \frac{1}{n}\sum_{h=1}^{n} r_{mh}\quad (h=1,2,\cdots,n) \tag{8}$$

The weight coefficient matrix $\boldsymbol{A}$ of correlation coefficient is expressed by formula (9).

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_m \end{pmatrix} \tag{9}$$

According to Table 1, the fuzzy relationship matrix $\boldsymbol{R}$ is built by following formula.

$$R = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ R_{m1} & R_{m2} & \cdots & R_{mn} \end{pmatrix} \tag{10}$$

The calculation of matrix is finished by formula (11).

$$B = A \circ R = \begin{pmatrix} A_1 & A_2 & \cdots & A_m \end{pmatrix} \circ \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ R_{m1} & R_{m2} & \cdots & R_{mn} \end{pmatrix} = \begin{pmatrix} B_1 & B_2 & \cdots & B_n \end{pmatrix} \tag{11}$$

With the principle of maximal membership, the result of printed character identification is got by the maximal fuzzy correlation measure, as showed in formula (12).

$$B_{\max} = \max\begin{pmatrix} B_1 & B_2 & \cdots & B_n \end{pmatrix} \tag{12}$$

6. Printing press identification experiment

16 printing presses are involved in one case. They are listed at Table 2.

Table 2 The type of 16 printing presses

No.	Type	No.	Type
1	HP Indigo press 5000	9	LD5100
2	SASY-600	10	DC1255/1250
3	CY47	11	PZ4660
4	JSASY-A	12	TruePress
5	JNYA-FCD	13	Nipson
6	HP Indigo 3050	14	Q-Press
7	DICO	15	DY452 II
8	TWS-212TEL	16	WS806MZ

6.1. The primary identification based on the minimal distance product of pixel distribution probability

According to the order of the minimal distance product (the value of Q), 6 printing presses are selected. Their values of Q are ordered, as showed in Table 3.

Table 3 The new sequence of 6 printing presses selected by distance product

No.	Type	No.	Type
B_1	TWS-212TEL	B_4	HP Indigo press 5000
B_2	DY452 II	B_5	LD5100
B_3	JNYA-FCD	B_6	JSASY-A

6.2. The ultimate identification based on the maximal fuzzy correlation measure of pixel distribution probability

(1) Correlation coefficient of printed characters

According to formula (5), the correlation coefficient of 6 printing presses can be got, as showed in Table 4.

Table 4 The distribution of correlation coefficient of 6 printing presses

	B_1	B_2	B_3	B_4	B_5	B_6
A_1	0.954	0.951	0.958	0.978	0.963	0.968
A_2	0.971	0.962	0.961	0.969	0.967	0.957
A_3	0.957	0.954	0.966	0.983	0.965	0.952

(2) The weight coefficient of segmentation mode

According to formula (7) and (8), the weight coefficient of segmentation mode of printed character can be calculated, as showed in Table 5.

Table 5 The weight coefficient of three segmentation modes

	A_1	A_2	A_3
Weight coefficient	0.339	0.304	0.357

(3) Printing press fuzzy identification

According to formula (10) to (12), Table 4 and 5, the result of printing press identification can be got.

$$\begin{aligned} B &= A \circ R \\ &= (0.339 \;\; 0.304 \;\; 0.357) \circ \begin{pmatrix} 0.954 & 0.951 & 0.958 & 0.978 & 0.963 & 0.968 \\ 0.971 & 0.962 & 0.961 & 0.969 & 0.967 & 0.957 \\ 0.957 & 0.954 & 0.966 & 0.983 & 0.965 & 0.952 \end{pmatrix} \\ &= (0.961 \;\; 0.955 \;\; 0.962 \;\; 0.977 \;\; 0.965 \;\; 0.959) \end{aligned} \quad (13)$$

$$B_{\max} = \max(B_1 \;\; B_2 \;\; B_3 \;\; B_4 \;\; B_5 \;\; B_6) = B_4 = 0.977 \quad (14)$$

This result is same with the result of identification of the minimal distance product. Therefore, the type of identified printing press is HP Indigo press 5000. The conclusion is accordant with the conclusion of this case. With 143 printing presses identification experiment, the accurate identification rate is about 96.18%.

7. Conclusion

The pixel distribution probability reflects the structural and statistic feature of printed Chinese character very well. With multi-segmentation mode and its interaction, the tiny difference of indexes is amplified markedly by the minimal distance product of pixel distribution probability of character image. On this condition, the double-optimized calculation of the minimal distance product and the maximal fuzzy correlation measure can realize the identification of printing press. The experiment showed that printing press identification based on the minimal distance product and the maximal fuzzy correlation measure of pixel distribution probability is simple, accurate, efficient and feasible.

Acknowledgment

It is a project supported by the Scientific Research Foundation of Guangzhou Public Security Bureau.

References

[1] Chen Li, Ding Xiaoqing, "Font Recognition of Single Chinese Character using stroke features", *Pattern Recognition and Artificial Intelligence*, 2004,Vol. 17, No.2, pp. 212-217.

[2] S. Khoubyari and J. Hul, "Font and Function Word Identification in Document Recognition", *Computer Vision and Image Understanding*, 1996, Vol. 63, No.1, pp. 66-74.

[3] Sergios Theodoridis, Konstantinos Koutroumbas, *Pattern Recognition* (*Second Edition*), Publishing House of Electronics Industry, Beijing, 2004(in Chinese).

[4] Ning Wang, "Chinese Character Fuzzy Recognition based on Strokes Distribution Probability of Three-Mode and Nine-Section", *Proceedings of The International Conference 2007 on Information Computing and Automation*. Chengdu, China, 2007, Vol. 3, pp.1170-1173.

[5] Lu Zong-qi, Jin Deng-nan, *Visual C++.NET Image Processing Programme*, Tsinghua University Publishing Company, Beijing, 2006(in Chinese).

[6] Yang Zhi-ling, Wang Kai, *Visual C++ Digital Image Acquisition, Processing and Application*, Post & Telecommunications Press, Beijing, 2003(in Chinese).

[7] Lu Y., "Machine Printed Character Segmentation-An Overview", *Pattern Recognition*, 1995, Vol. 28, No.1, pp. 67-80.

[8] Wang H W, Ma G F, and Wang Z C, "The Study of Fuzzy Identification Theory and Its Practical Applications", *Journal of System Simulation*, 2000, Vol. 12, No. 3, pp. 87-90.

[9] *Mathematics Handbook*, China Coal Industry Publishing House, Beijing, 1976(in Chinese).

[10] He Z X, *Fuzzy Mathematics and Its Application*, Tianjin Science and Technology Press, Tianjin, 1983(in Chinese).