Single character type identification

Yefeng Zheng*, Changsong Liu, Xiaoqing Ding

Department of Electronic Engineering, Tsinghua University
Beijing 100084, P.R. China

ABSTRACT

Different character recognition problems have their own specific characteristics. The state-of-art OCR technologies take different recognition approaches, which are most effective, to recognize different types of characters. How to identify character type automatically, then use specific recognition engines, has not brought enough attention among researchers. Most of the limited researches are based on the whole document image, a block of text or a text line. This paper addresses the problem of character type identification independent of its content, including handwritten/printed Chinese character identification, and printed Chinese/English character identification, based on only one character. Exploiting some effective features, such as run-lengths histogram features and stroke density histogram features, we have got very promising result. The identification correct rate is higher than 98% in our experiments.

Keywords: character type identification, handwritten/print identification, Chinese/English identification, script identification

1. INTRODUCTION

Different character recognition problems have their own specific characteristics. The state-of-art OCR technologies take different approaches, which are most effective, to recognize characters with different scripts. For example, different features and classifiers are exploited to recognize English and Chinese characters. Even within the same script, handwritten and printed characters are different with each other drastically. We take different techniques to recognize them too. In most out-of-shelf commercial OCR products, users are needed to indicate the character type of a text block. Then, the OCR products call different recognition engines. How to identify character type automatically, especially single character type identification, has not brought enough attention among researchers.

In handwritten/printed identification, most techniques in references are based on one or several text lines[1,2]. Printed text lines are arranged regularly, while handwritten text lines are arranged with large variance. K. C. Fan[2] made use of this distinction to identify handwritten/printed text lines. The merit of this approach is that it can be used in different script. For example, the scheme for handwritten/printed Chinese text lines identification can be used to identify English handwritten/printed text lines without or with little modification. However it only makes use of inter-character property, and not use intra-character property, which is very useful to identify handwritten/printed characters. Therefore identification correct rate is not very promising, only about 85%[2].

---

*Yefeng Zheng is now pursuing PhD at the University of Maryland, College Park.
He can be reached at zhengyf@cfar.umd.edu

Due to the increasing demand of international communication, multi-script OCR attracts more and more attention. There are more researches in script identification than in handwritten/printed identification. Most techniques in references are based on the whole document image[3], several text lines[4] or a text block[5]. However, in some cases, different scripts are mixed together. For example, English characters are used together with Chinese characters in typical technical reports nowadays in China. Furthermore, sometimes, we cannot get a whole text line. For example, in form recognition, many form cells have only 1 to 3 characters, such as the cells filled with name, sex, nation, etc. So in the above situations, we should do script identification on single character. This paper presented a method to identify single character type, including handwritten/printed Chinese identification and printed Chinese/English identification.

2. CHARACTER STROKE WIDTH NORMALIZATION

The stroke width of different fonts varies greatly, as shown in Figure 5. Such large variance will bring some problems for us to get correct identification. So stroke width normalization is indispensable in preprocessing stage. Our normalization approach is a modified version of M. D. Garris' method[6], which is a kind of mathematic morphology approach. The basic idea is: if the stroke is too thick (e.g. larger than 3 pixels), do erosion operation; if the stroke is too thin (e.g. less than 2 pixels), do dilation operation. However, some Chinese fonts have different stroke width for horizontal and vertical strokes. In such fonts, normally, the horizontal stroke is very thin, only 1 to 2 pixels, while vertical stroke is very thick, as shown by the left character in figure 1. If we do the same erosion operation to horizontal and vertical strokes, some horizontal strokes will be eroded by error, as shown by the middle character in figure 1. To solve this problem, horizontal and vertical strokes should be processed differently. First, the horizontal and vertical stroke width is estimated using run length histograms. If there are no remarkable different between the horizontal and vertical stroke width, we process them using the same morphologic operation. Otherwise, when we erode or dilate a pixel, we must decide which type of stroke it belongs to. We can judge whether the pixel belongs to a horizontal or a vertical stroke by extracting the horizontal and vertical run lengths passing through this pixel. If the horizontal run length is longer than the vertical one, then the pixel belongs to a horizontal stroke. Otherwise, it belongs to a vertical stroke. Different morphologic operations are used to process pixels of horizontal and vertical strokes. Figure 1 shows the character width normalization result. The left is the character image before normalization. The middle is the character after erosion using M. D. Garris' method. The right is the erosion result using the modified approach. The horizontal strokes do not change after normalization, still 2 pixels width, while vertical strokes are eroded by 1 pixel.

Figure 1. Character width normalization

3. HANDWRITTEN/PRINTED CHINESE CHARACTER IDENTIFICATION

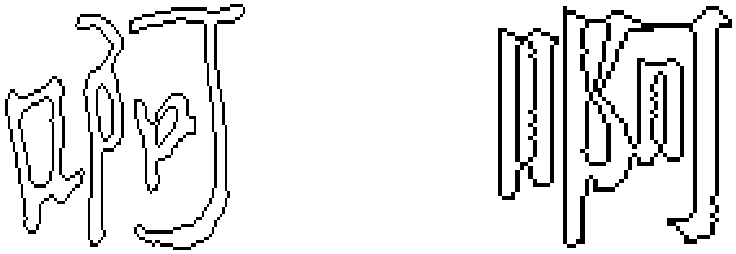
Chinese characters are made up of four kinds of strokes: horizontal, vertical, Pie and Na. The latter two kinds of strokes are skewed about 30 to 60 degrees and –30 to –60 degrees respectively. About 75% of Chinese character strokes are horizontal and vertical. In the printed Chinese characters, the horizontal and vertical strokes are straight, while those strokes in handwritten Chinese characters are cursive. This is the most distinctive difference between handwritten and printed Chinese

characters. Represented in the run length histogram, printed characters have more long run lengths than handwritten characters, as shown in Figure 2. We divide the histogram into five bins equally. Taking the number of run lengths in each bin as features, we get 20 features from four histograms according to horizontal, vertical, 45 degree and -45 degree run lengths.

(a) (b)

Figure 2. (a) Vertical run length histogram of handwritten Chinese character 啊
(b) Vertical run length histogram of printed Chinese character 啊

Run length histogram features are global, cannot capture the local deform of handwritten characters. So we add local features: weighted chain code features. Figure 3 shows the contours of handwritten and printed Chinese character 啊. On the printed character contours, the vertical strokes are very straight and made up with some continued same chain code. While the chain codes of the contours of handwritten characters are interrupted by different chain codes from time to time due to the local deform of handwriting. To capture this distinction, four attributes are given to every contour point, according to horizontal, vertical, 45 degree and -45 degree chain codes. Taking horizontal attribute for example, the horizontal attribute of a contour point is the number of continuing horizontal chain code passing through this point. Adding up the attributes of all contour points, we get four weighted chain code features.



(a) (b)

Figure 3. (a) Contour of handwritten Chinese character 啊
(b) Contour of printed Chinese character 啊

Finally, including aspect ratio, we get 25 features to identify handwritten/printed Chinese characters.

4. PRINTED CHINESE/ENGLISH CHARACTER IDENTIFICATION

There are four major differences among Chinese and English (including digits) characters:

1). Chinese characters are more complex in topology. About 60% Chinese character have two or more than two connected components. However, if we filter the small dots on the top of the characters “i” and “j”, all English characters have only one connected component.

2). The holes of English characters (including digits) and their arrangement are very regular. There are only three situations: first, no holes, e.g. “C”; second, one hole, then the hole is located on the center of the character, e.g. “A” and “D”; third, two holes, then the two holes are arranged vertically, e.g. “B” and “8”.

3). There are more horizontal and vertical strokes than cursive strokes in Chinese characters. As mentioned in the above section, about 75% of the strokes of Chinese characters are horizontal and vertical. While, low case English characters have more cursive strokes. This distinction can be captured by run length histogram features discussed above.

4). Chinese characters have more strokes than English characters. Their stroke densities are different. Figure 4 shows the stroke density of a Chinese character. The scan line penetrates the character five times, so the vertical stroke density of the line is five. Then, we can get the histogram of the horizontal and vertical stroke density respectively. Similar to run length histogram features, we divide the histogram into five equal wide bins, and then take the line number in each bin as the stroke density histogram features.

Finally, we use aspect ratio, black pixel density, connected component number, hole number, run length histogram features and stroke density histogram features, to identify printed Chinese/English character.

Figure 4. Stroke density

If the quality of character images is good, using connected components, holes and their arrangement, about 80% Chinese characters can be identified correctly. The other Chinese characters can be identified from English characters using other features discussed above. However, with poor image quality, about 1.7% English character samples are identified as Chinese characters by error, due to broken and noise, etc. So we use these two features together with other features to get statistical classification.

5. CLASSIFIER DESIGN

There are thousands of Chinese characters, among which about 3755 characters are frequently used. Furthermore, Chinese characters vary greatly in both stroke arrangement and complexity. For example, some Chinese characters only have horizontal strokes, while some Chinese character have no horizontal and vertical strokes at all. Some characters are very simple with only one or two strokes, while some are very complex with more than twenty strokes. Therefore, the extracted features are greatly influenced by the variance of Chinese characters. The within-class difference of run length histogram features is very large. Using BP neural network, we cannot get good result, either no convergence or over fitting. The identification correct rate is even worse than Fisher classifier.

Due to the powerful features we used, with Fisher classifier, we get very promising result. The identification correct rate is as high as 96%. SVM classifier is used to increase the correct rate further. SVM classifier is based on the VC dimension

theory and structural risk minimum theory of statistical learning[7]. It has no annoying local maximum problem of the neural network. So it is much easier to train a SVM classifier. In our experiment, using SVM classifier, the correct rate is higher than 98%.

However, the classification speed of SVM is rather slow. It can be shown from the classification formula of SVM [7]:

$$f(x) = sign(\sum_{i=1}^{l} y_i a_i K(x_i, x) + b) \quad (1) \qquad (1)$$

Here, y_i if the class label of support vector x_i, taking value 1 or -1; a_i is the weight of the support vector x_i; $K(x_i, x)$ is the kernel function. In our experiment it is Gauss function; b is a constant; l is the number of support vectors. The classification speed of SVM is determined by the number of support vectors, which has close relation to the Bayes error[7]. In our experiment the support vector number is about 1000~2000. To make a decision, l times Gauss function calculations and one vector multiplication are needed. While only one vector multiplication is needed to make a decision in Fisher classifier. The speed of Fisher classifier is faster than SVM classifier by several magnitudes.

We try to combine these two classifiers, getting a tradeoff between high speed of Fisher classifier and the high correct rate of SVM classifier. First calculate the Fisher classification confidence. In Fisher classifier, the feature vector is projected onto an axis, on which the ratio of inter-class diversity to inter-class diversity is maximized. Due to the central limit theorem, the distribution of the projection can be approximated by a Gauss distribution.

$$f_p(y) = \frac{1}{\sqrt{2p}\, s_p} \exp[-\frac{1}{2}(\frac{y - m_p}{s_p})^2] \qquad (2)$$

$$f_h(y) = \frac{1}{\sqrt{2p}\, s_h} \exp[-\frac{1}{2}(\frac{y - m_h}{s_h})^2] \qquad (3)$$

$f_p(y)$ and $f_h(y)$ are the probability density functions of the projection of printed Chinese characters and handwritten Chinese characters respectively. The parameters, m_h, m_p, s_h, s_p, are estimated from the training samples. So the confidence, which is the posteriori probability of the decision, is:

$$Conf = \begin{cases} \dfrac{f_h(y)}{f_p(y) + f_h(y)} & if\ f_h(y) > f_p(y)\quad Classified\quad as\quad handwritten \\ \dfrac{f_p(y)}{f_p(y) + f_h(y)} & if\ f_p(y) \geq f_h(y)\quad Classified\quad as\quad printed \end{cases} \qquad (4)$$

If the Fisher classification confidence is higher than a threshold (e.g. 90%), then the classification result is accepted directly. Otherwise, SVM classifier is used to do further classification. By adjusting the threshold, we can get a tradeoff between speed and accuracy dynamically.

6. EXPERIMENTS

6.1 Handwritten/Printed Chinese Character Identification

Total 375,487 printed Chinese character samples and 371,186 handwritten Chinese character samples are used in our experiment, including 3755 different characters. Parts of the samples are shown in figure 5. The first row is handwritten Chinese characters, the second row is printed Chinese characters, and the last row is printed English characters, which are used for the next experiment. Randomly taking about 20,000 samples from each class as training set. The remained samples used as test set. The experiment result is shown in table 1.

Using Fisher classifier, on test set, the identification correct rate of handwritten Chinese characters is as high as 96.0%, the correct rate of printed Chinese characters is 96.5%. Using SVM classifier, we can get more promising result: 98.2% for handwritten Chinese characters and 99.1% for printed Chinese characters.

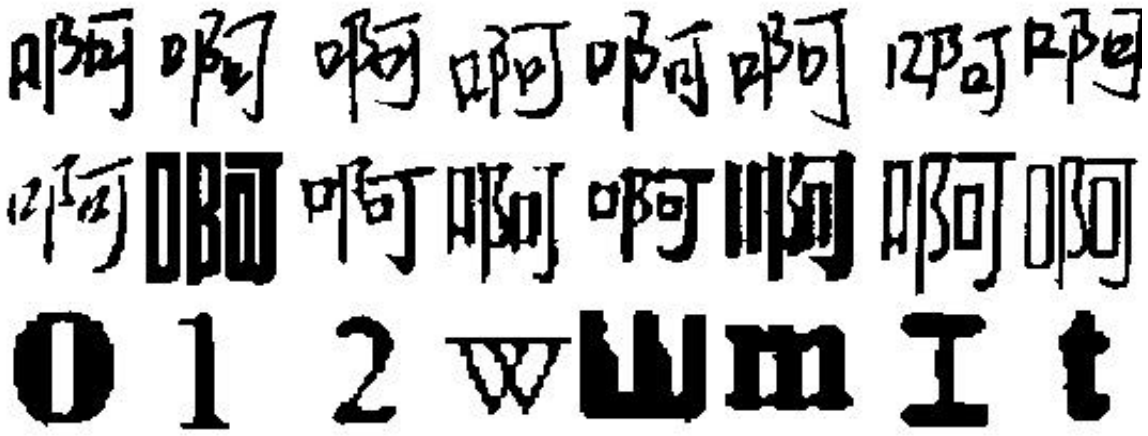


Figure 5. Parts of experiment samples

On a PC with 500 MHZ PII CPU, the speed of feature extraction is about 1.4 ms/character. The classification speed of Fisher classifier is very fast, processing more than 10^7 characters per second. The computation demand is negligible compared to the feature extraction. However the classification speed of SVM classifier is rather slow, about 3.2 ms/character. Adding the feature extraction, the total identification speed using Fisher classifier is 720 characters/second, and that of SVM classifier is 220 characters/second. We combined these two classifiers together. First using Fisher classifier, those training samples with identification confidence less than 85% are used to train the SVM classifier. There are about 5,000 samples used to train the SVM classifier. In identification stage, the identification confidence higher than 85% is accepted directly after Fisher classification. From table 1, we can see that about 86% characters get result after Fisher classification. The remained samples are put into the SVM classifier. By setting different threshold, we can get tradeoff between speed and accuracy dynamically. Using 85% as the threshold, the total identification speed is about 560 characters/ second. After combination, compared to Fisher classifier, the speed does not decrease too much, however the error rate decreased by 50%. Compared to SVM classifier, the speed increased to 2.5 times, while the accuracy decreased only by 0.4%.

Using Fisher classifier, the top 10 most error prone handwritten Chinese characters are: 二一口工土曰田三入八. The error rates of these characters range from 50% to 90%. The errors of top 10 characters are 4.5% of the total errors of handwritten Chinese characters. The top 10 most error prone printed Chinese characters are: 淡乡夕诊沙灸广汐济多. The error rates of these characters range from 40% to 60%. The errors of top 10 characters are about 3.8% of the total errors of

printed Chinese characters. From the above result, we can see that 8 of the top 10 most error prone handwritten Chinese character only have horizontal and vertical strokes. However, the top 10 most error prone printed Chinese characters have few horizontal and vertical strokes. Some print fonts are very similar to handwritten type, as shown by the second row first column character in figure 5. They will be identified as handwritten characters with high probability.

Table 1. Handwritten/printed Chinese Character Identification Result

		Training set		Test set	
		Handwritten	Printed	Handwritten	Printed
Fisher classifier	Correct rate	96.2%	96.3%	96.0%	96.5%
SVM classifier	Correct rate	98.7%	99.4%	98.2%	99.1%
First Fisher classifier (Threshold 85%)	Correct rate	85.2%	87.5%	85.4%	87.3%
	Error rate	1.1%	0.6%	1.1%	0.6%
	Rejection rate	13.7%	11.9%	13.5%	12.1%
Second SVM classifier	Correct rate	93.4%	95.6%	92.0%	94.6%
Total result of Fisher + SVM	Correct rate	98.0%	98.8%	97.8%	98.7%

6.2 Printed Chinese/English Character Identification

The printed Chinese character samples are the same to the above experiment. Printed English character samples are collected from real documents with poor quality. Total number of English samples is 422,000. Randomly taking about 20,000 samples from each class as training set. The remained samples used as test set. The identification result is shown in table 2. Using Fisher classifier, on test set, the identification correct rate of Chinese characters is 97.1%, and that of English characters is 96.6%. Using SVM classifier, the accuracy increased, the identification correct rate of Chinese and English characters both increased to 99.3%.

Table 2. Printed Chinese/English Character Identification Result

		Training set		Test set	
		Chinese	English	Chinese	English
Fisher classifier	Correct rate	97.0%	96.7%	97.1%	96.6%
SVM classifier	Correct rate	99.7%	99.6%	99.3%	99.3%
First Fisher classifier (Threshold 90%)	Correct rate	89.5%	86.9%	89.3%	86.8%
	Error rate	0.3%	0.9%	0.4%	0.9%
	Rejection rate	10.2%	12.2%	10.3%	12.3%
Second SVM classifier	Correct rate	98.6%	97.6%	95.9%	95.7%
Total result of Fisher + SVM	Correct rate	99.6%	98.9%	99.2%	98.6%

Using Fisher classifier, the top 10 most error prone Chinese characters are: 卜丫入人十广上下卞犬. The error rates of these characters range from 94% to 100%. These characters are almost identified by error every time. The errors of the top 10 Chinese characters are 9.3% of the total errors of Chinese characters. The top 10 most error prone English characters, including digits, are: “mWZgMw8BQX”. The error rates of these characters range from 7% to 45%. The errors of top 10 characters are 72% of the total errors of English characters. Compared to handwritten/printed Chinese character identification, we can see that most errors are concentrated on a few characters for both Chinese and English characters. Some Chinese characters are very similar to English characters, e.g. “K” and “卜”, “Y” and “丫”, “W” and “山”, “M” and ”爪”, ”D” and ”口”, “T” and “十” etc. So the errors are concentrated on these characters.

On the same PC as above, the speed of feature extraction is 1.2 ms/character. Taking 90% as the threshold, the total speed of Fisher + SVM classifier is about 650 characters/second.

7. CONCLUSION

Most of the existing character type identification algorithms are based on a whole document image, a text block, or a text line. We focused our research on single character type identification, and have got very promising result.

The further research on handwritten/printed Chinese character identification includes finding more powerful local features to capture the local distortion of handwritten Chinese characters. There are thousands of Chinese characters with great variance. The features used in our experiments are influenced by the diversity of characters greatly, which bring the problem of large within-class variance. One possible approach to solve this problem is to separate the characters into several clusters. For examples, we can get three clusters, representing Chinese characters with simple, middle and complex strokes respectively. Then we can train and use three different classifiers. By splitting a complex pattern recognition problem into several small and simple problems, we can get higher accuracy. This approach is valid to printed Chinese/English character identification too. How to solve the problem of intrinsic similarity between some Chinese and English characters needs more study.

8. ACKNOWLEDGEMENTS

We used LIBSVM, developed by C-C. Chang and C-J. Lin, for our support vector machine classification. And Miss Peng of our lab gave us a lot of help to prepare samples for experiments.

REFERENCES

1. S. N. Srihari, Y. C. Shim, and V. Ramanaprasad, "A system to read names and address on tax forms", *Technical Report CEDAR-TR-94-2*, CEDAR, SUNY, Buffalo, 1994
2. K. C. Fan, L. S. Wang, and Y. T. Tu, "Classification of machine-printed and handwritten texts using character block layout variance", *Pattern Recognition, 31(9),* pp: 1275-1284, 1998
3. J Hochberg, P. Kelly, and T. Thomas, "Automatic script identification from document images using cluster based templates", *IEEE Trans. On Pattern Analysis & Machine Intelligence, 19(2),* pp: 176-182, 1997
4. A. L. Spitz, "Determination of the script and language content of document images", *IEEE Trans. On Pattern Analysis & Machine Intelligence, 19(3),* pp: 235-245, 1997
5. T. N. Tan, "Rotation invariant texture features and their use in automatic script identification", *IEEE Trans. On Pattern Analysis & Machine Intelligence, 20(7)*, pp: 751-756, 1998
6. M.D. Garris, J.L. Blue, and G.T. Candela etc., "NIST Form-Based Handprint Recognition System", *Internal Report 5469 and CD-ROM*, National Institute of Standards and Technology, July 1994.
7. C. Cortes, V. Vapnik, "Support-Vector networks", *Machine Learning, 20*, pp: 273-297, 1995