International Journal of Pattern Recognition and Artificial Intelligence
Vol. 16, No. 1 (2002) 85–96

SEGMENTATION OF HANDWRITTEN CHINESE CHARACTERS FROM DESTINATION ADDRESSES OF MAIL PIECES

YUE LU* and CHEW LIM TAN

Department of Computer Science, School of Computing, National University of Singapore, Singapore 119260
**luy@comp.nus.edu.sg*

PENGFEI SHI and KEHUA ZHANG

**Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200030, China*

In this paper, we illustrate a method to segment handwritten Chinese characters from destination addresses of mail pieces. Fast Hough transform is utilized to detect the reference lines preprinted on the mail piece. In the segmentation, subassemblies of Chinese characters are merged based on the structural features of Chinese characters and the subassemblies' topological relations, viz. upper–lower, inside–outside and left–right relations. The width of subassemblies and the spacing between neighboring subassemblies in the whole image of the destination address are analyzed to guide the merging of the left–right subassemblies. Experimental results with real mail piece images show that the proposed approach has achieved a promising performance for segmenting handwritten Chinese characters.

Keywords: Document image processing; postal automation; handwritten Chinese character; character segmentation; detection and removal of reference lines.

1. Introduction

Automatic letter sorting is a key section in the postal automation, in which character recognition technology has been successfully applied. In recent years, with the rapid development of Chinese economy, more than 200 sets of automatic letter sorters, by which the postal numerals written in the six boxes can be recognized, have been deployed throughout China so far. However, it is hard to significantly improve the performance of numeral recognition based on the current theory and technology in the field of character recognition. On the other hand, the information included in the destination address of a mail piece is generally ignored. So it is of significant meaning to study a practical algorithm for segmenting and recognizing Chinese characters in the destination address.

As a matter of fact, segmentation is a very important procedure for offline handwritten Chinese character recognition system, because most recognizers can only deal with isolated characters and correct recognition relies on correct

segmentation of characters. Therefore, segmentation of handwritten Chinese characters from document images plays a crucial role in the recognition system.

A number of papers concerning the offline recognition of handwritten Chinese characters have been published in the past several decades,[12,14,16] and almost all of them focus on the recognition of isolated characters. However, the segmentation of free-format handwritten Chinese characters from document images has been rarely discussed except in Refs. 6 and 11. In Hong's approach,[6] the basic segmentation and the fine segmentation were performed based on the varying spacing thresholds and the minimum variance criteria. The five most probable ways of segmentation were derived, and all the possible segments were extracted and recognized. Then a lattice was created and searched using a viterbi-based algorithm to find the most likely character sequence. This method was proposed on the assumption that the gap consequentially existed between two neighboring characters; therefore it cannot deal with the overlapping characters.

Tseng and Chen[11] presented an approach based on heuristic merging of stroke bounding boxes and dynamic programming to segment handwritten Chinese characters. The strokes of the Chinese characters were extracted first and were used to build the stroke bounding boxes. Knowledge-based merging operations were applied to merge the stroke bounding boxes to form candidate boxes. Finally, a dynamic programming method was utilized to find the best segmentation boundaries. This method was able to handle the overlapping of characters, but the extraction of character strokes is time-consuming.

Several methods have been reported for the segmentation of handwritten western alphabets and numerals.[2,7,9,10] Most of these approaches make use of spatial information or character shape information. In general, the segmentation can be classified into two strategies. One is the structure-based technique, and the other is the recognition-based technique. However, Chinese characters are quite different from the western alphabets and numerals. Their structure is more complicated. Most Chinese characters consist of more than two connected components, and the spacing between characters and the spacing inside characters vary considerably. Though it is common for people to leave a larger gap between characters than inside characters, the spatial information is not reliable.

This paper will focus on segmenting handwritten Chinese characters from destination addresses of mail pieces. In the next section, we describe the preprocessing in which the reference line preprinted on a mail piece face is detected and removed from the image of the destination address region. While Hough transform is utilized to detect the reference line, the extracted feature points are divided into two subsets, from which two feature points are selected respectively to compute parameters in the Hough space.

The segmentation of handwritten Chinese characters is discussed in Sec. 3. According to the structural features of Chinese characters, subassemblies are merged based on their topological relations, viz. upper–lower, left–right and inside–outside. The width of subassemblies and spacing between neighboring subassemblies in

the whole image of the destination address are analyzed to guide the merging of left–right subassemblies.

Section 4 gives the test results of the proposed approach which is applied to segment the handwritten Chinese characters from the destination addresses of real mail pieces.

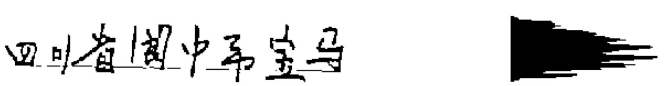
2. Detection and Removal of the Reference Line

In general, three horizontal straight lines are preprinted on many Chinese envelopes to standardize the writing habit. These lines correspond to the destination address line (DAL), the addressee name line (ANL) and the sender address line (SAL), respectively. These are called reference lines. Consequently, detection and removal of reference lines play a very important role in the preprocessing.

There are two typical methods of line detection. One is based on projection profile, and the other one is on Hough transform. The former is a time-tested technique, but makes the implicit assumption that the line was preprinted horizontally and that the image was captured without skew. Hence, the method fails for lines with significant tilt. Unfortunately, when an envelope is fed mechanically to a letter sorting machine, it suffers from some degrees of skew or tilt. A few degrees skew is unavoidable. It is a fact that the reference line cannot be detected successfully by this approach in many cases, as illustrated in Fig. 1.

Hough transform[8] is a commonly used line detecting method, and has been widely applied in the document skew detection.[1,5] The basic idea of straight line detection by the standard Hough transform is a voting process where each pixel on the image votes for all the possible lines passing through that point. Each pixel on the image space is transformed to a parameterized line in the parameter space. An accumulator with a cell array is laid on the parameter space, and each image pixel gives one score to the cell lying on its parameterized line. The parameter coordinates of the cell with the maximal score are used to represent the straight line of the image space. The heavy burden of the computational complexity and massive storage requirement is the primary drawback of standard Hough transform. Besides, it is also quite a blind algorithm. To overcome these weaknesses, many modified approaches have been presented.[3,13]

To meet the real-time requirements of the practical system, the knowledge of the reference line on the mail piece image is utilized to alleviate the computational and storage burden of Hough transform in this paper. In summary, some improvements



(a) (b)

Fig. 1. Reference line cannot be detected by horizontal projection. (a) Original image; (b) projection.

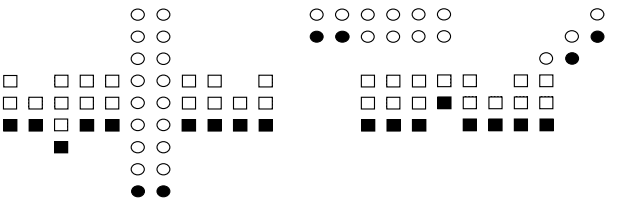


Fig. 2. Selection of feature points. • Pixels of strokes, and feature points; ○ Pixels of strokes, but not feature points; ■ Pixels of reference line, and feature points; □ Pixels of reference line, but not feature points.

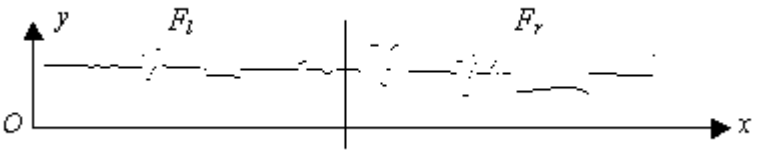


Fig. 3. Feature set of Fig. 1.

are made:

(1) For a reference line expressed by $y = ax+b$, it is generally located within a small angle with the horizon. In a heuristic way, the parameter a can be set between $[-A, A]$, which can reduce the search range significantly. The parameter A is equal to $\arctan(H_{\text{dai}}/W_{\text{dai}})$, where H_{dai} and W_{dai} represent the height and width of the destination address image, respectively.

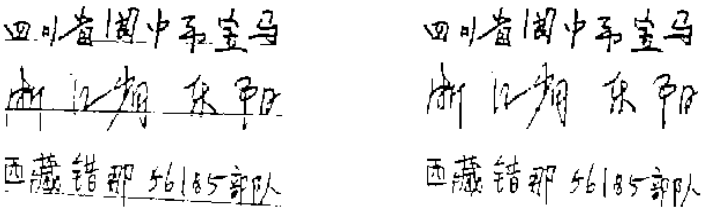
(2) Only feature points are transformed to the parameter space instead of all pixels. The reference line is located at the bottom of the destination address region. The set F of the feature points consists of the lowest pixels in each column. Although some character strokes may cross the reference line, the feature points mostly come from the underside edge of the reference line, as illustrated in Fig. 2. This results in a drastic reduction in the computation time. Figure 3 shows the feature points corresponding to Fig. 1.

(3) The features are divided into two subsets: F_l and F_r. They correspond to the left and the right parts of the destination address image. Obtaining two feature points $p_1(x_1, y_1) \in F_l$ and $p_2(x_2, y_2) \in F_r$, we can simply map them into a point (a, b) of the parameter space by solving the following joint equations:

$$\begin{cases} y_1 = ax_1 + b \\ y_2 = ax_2 + b \end{cases}$$

Viz.

$$\begin{cases} a = \dfrac{y_2 - y_1}{x_2 - x_1} \\ b = \dfrac{x_2 y_1 - x_1 y_2}{x_2 - x_1} \end{cases}$$



(a) (b)

Fig. 4. Detection and removal of reference line on destination address. (a) Original images; (b) results of removing reference lines.

In view of the fact that x_2 is not equal to x_1, the equations can be solved for the correct a and b. In this way, the blind search and high computation cost in the conventional Hough transform are avoided.

(4) Once there is an accumulated cell with a score larger than the threshold N, the detection procedure is stopped, and the corresponding parameter is considered as the detected reference line. The threshold N is related to the width of destination address image. We set $N = \tau W_{\text{dai}}$, where τ is a constant factor.

The detected reference line is removed according to the local position relation between the reference line and character strokes, such as contact, intersection and superposition, as discussed in our earlier paper.[15] Figure 4 demonstrates the effectiveness of the method by experiments on detection and removal of reference lines on the real Chinese mail pieces. The algorithm is run on a PII-300. The detection of reference line by Hough transform requires 4–11 milliseconds for one mail piece, and the removal of reference line requires about 3 milliseconds on the average. It can be seen that the practical value can be achieved by the present approach.

3. Segmentation of Chinese Characters

3.1. *Characteristic of Chinese characters*

The structure of Chinese characters is more complex than that of the western characters. Most Chinese characters consist of more than two connected components.[4]

For convenience, we define the *subassembly* of a Chinese character in this paper.

Definition 1. There are n connected components $C^{(k)}(k = 1, 2, \ldots, n)$ in the image of a given Chinese character H, which is denoted as $H = \bigcup_{\forall\, k\in\{1,2,\ldots,n\}} C^{(k)}$, then the *subassembly* E is defined as $E = \bigcup_{k\in\{1,2,\ldots,n\}} C^{(k)}$, which means that E is a composition of one or several connected components.

The subassembly acts as a bridge between the connected component and the complete Chinese character, where the subassembly is no less than one connected component, but not bigger than a complete Chinese character. In general, a subassembly includes one or several connected components, and a Chinese character includes one or several subassemblies.

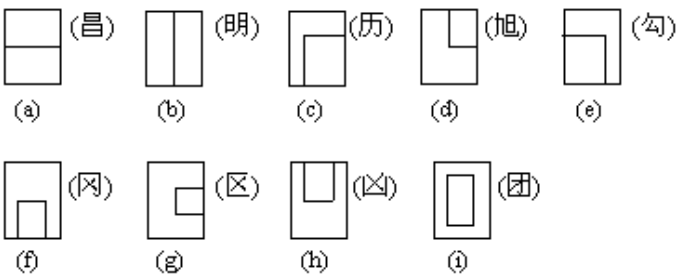


Fig. 5. Topological relations between two subassemblies.

It is found that the topological relation of any two assemblies can be expressed as one of the nine relations showed in Fig. 5.

3.2. *Merging of subassemblies*

The topological relation of any two subassemblies belongs to one of the three types, viz. upper–lower [Fig. 5(a)], left–right [Fig. 5(b)] and inside–outside [Figs. 5(c)–5(i)]. Different processings are applied on different types as follows:

(a) upper–lower: merge into one Chinese character.
(b) inside–outside: merge into one Chinese character.
(c) left–right: merge according to the position information of neighbors.

Position information of each subassemblies (m) includes width $W^{(m)}$, height $H^{(m)}$, upper-left position $(LT_x^{(m)}, LT_y^{(m)})$, bottom-right position $(RB_x^{(m)}, RB_y^{(m)})$ and central position $(C_x^{(m)}, C_y^{(m)})$, where $1 \leq m \leq M$, M is the total number of connected components. The following operation will merge the subassembly (m) with the subassembly (n) to become subassembly (k):

$$\begin{aligned}
LT_x^{(k)} &= \min(LT_x^{(m)}, LT_x^{(n)}) \\
LT_y^{(k)} &= \min(LT_y^{(m)}, LTy_x^{(n)}) \\
RB_x^{(k)} &= \max(RB_x^{(m)}, RB_x^{(n)}) \\
RB_y^{(k)} &= \max(RB_y^{(m)}, RB_y^{(n)}) \\
W^{(k)} &= RB_x^{(k)} - LT_x^{(k)} \\
H^{(k)} &= RB_y^{(k)} - LT_y^{(k)} \\
C_x^{(k)} &= (RB_x^{(k)} + LT_x^{(k)})/2 \\
C_y^{(k)} &= (RB_y^{(k)} + LT_y^{(k)})/2
\end{aligned}$$

It is evident that the merging of upper–lower and inside–outside subassemblies is trivial. However, the free-format writing results in the difficulty of merging the left–right subassemblies. An analysis of the structure of Chinese characters and the writing style of the character string is helpful.

3.3. *Merging of left–right subassemblies*

The merging of left–right subassemblies is a challenging problem. Generally speaking, whether the neighboring subassemblies should be merged into one character depends on the information such as the width of subassemblies and the spacing between the neighboring subassemblies. However, as the writing style significantly varies from person to person, the above information is not stable. Analysis of the writing style of the whole destination address can provide heuristic information for merging the left–right subassemblies, which is based on the following hypothesis:

(a) In general, the given destination address was written by a certain writer at one time, and the writing style is somewhat consistent for a given image.
(b) Normally, the spacing between characters is wider than that between subassemblies of one character.
(c) The difference of sizes (width and height) among written Chinese characters in the destination address is not too big.

As illustrated in Fig. 6, only left–right subassemblies exist after the upper–lower and inside–outside subassemblies have been merged. Suppose that the spacing of the neighboring subassemblies is $G_e^{(k)}(k = 1, 2, \ldots, M-1)$, and $W_e^{(k)}(k = 1, 2, \ldots, M)$ is the width of each subassembly, where M is the number of subassemblies.

The width of the merged subassembly combining (k) and $(k+1)$ is to be

$$W_m^{(k)} = G_e^{(k)} + W_e^{(k)} + W_e^{(k+1)} \quad k = 1, \ldots, M-1$$

$G_e^{(k)}$ and $W_m^{(k)}$ are normalized as:

$$g_e^{(k)} = G_e^{(k)} / \max Ge$$

$$w_m^{(k)} = W_m^{(k)} / \max Wm$$

where

$$\max Ge = \max_{1 \le k \le M-1} G_e^{(k)}$$

$$\max Wm = \max_{1 \le k \le M-1} W_m^{(k)}$$

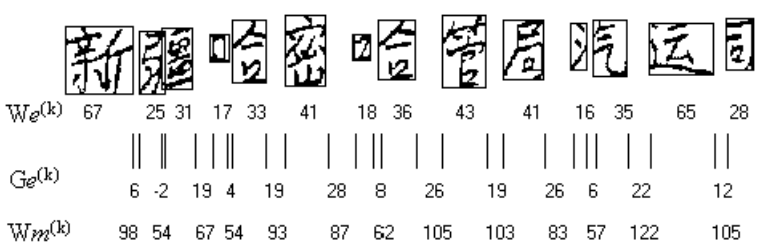


Fig. 6. The width of subassemblies and spacing between neighboring subassemblies.

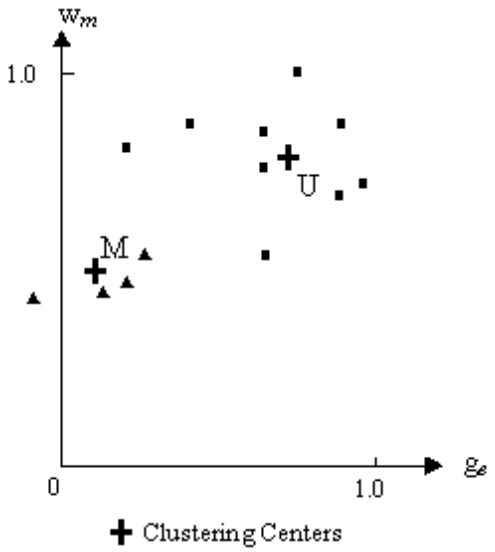


Fig. 7. Clustering centers generated by left–right assemblies.

$g_e^{(k)}$ and $w_m^{(k)} (k = 1, \ldots, M-1)$ are taken as the parameters describing the subassemblies pair $[(k), (k+1)]$. They become two clustering centers $M(C_g^m, C_w^m)$ and $U(C_g^u, C_w^u)$, respectively. If $[(k), (k+1)]$ is closer to the clustering center $M(C_g^m, C_w^m)$, the subassemblies pair are inclined to be merged. On the contrary, if the $[(k), (k+1)]$ is close to the clustering center $U(C_g^u, C_w^u)$, the subassemblies pair are not inclined to merge. Figure 7 is the clustering result of Fig. 6. This clustering will guild the merging of the left–right subassemblies.

Furthermore, the average width of Chinese characters W_a can be evaluated by the subassemblies that are closer to the clustering center $U(C_g^u, C_w^u)$. The average width W_a is also the heuristic information for merging the left–right subassemblies.

After all of the left–right subassemblies are ordered according to $C_x^{(k)}$ (see Sec. 3.2), the following algorithm is developed to merge them.

Algorithm of merging the left–right subassemblies

Input: position information of subassemblies according to monotonically increasing $C_x^{(k)}$.

Output: character bounding boxes.

Procedures:

Step 1. Initialize flag = F

Step 2. Calculate the clustering centers $M(C_g^m, C_w^m)$ and $U(C_g^u, C_w^u)$

Step 3. Evaluate W_a

Step 4. For all subassemblies k

If $(RB_x^{(k+1)} - LT_x^{(k)}) < \varepsilon W_a$ .and. $[(m),(m+1)]$ is closer to $M(C_g^m, C_w^m)$

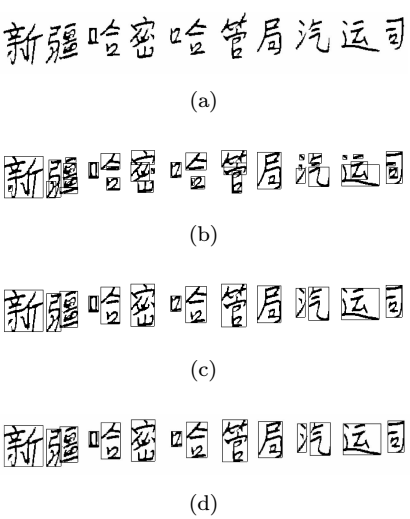
Then Merge (m) and $(m+1)$, flag $= T$

Step 5. If flag $\neq F$ go to step 1, else stop.

Where ε is a constant factor. We let $\varepsilon = 1.2$ experimentally.

3.4. *Experimental results*

The proposed segmentation approach is applied to segment the handwritten Chinese characters from destination addresses of mail pieces. The procedure has been demonstrated in Fig. 8. As the result of subassemblies merging, each handwritten Chinese character is enclosed successfully by a rectangle. Figure 9 gives some



(a)

(b)

(c)

(d)

Fig. 8. Segmentation procedure. (a) Original image; (b) connected components; (c) subassemblies after upper–lower and inside–outside merging; (d) segmented character after left–right merging.

Fig. 9. Segmentation results of some destination address images on the real mail pieces.

segmentation results of some destination address images captured from the real mail pieces.

4. Test Results and Conclusions

For the experiments, 3,498 images containing 26,131 handwritten Chinese characters were captured from real Chinese mails by an automatic mail sorting machine. Each destination address image was segmented from the destination address zone first, and then the experiments were conducted to segment handwritten Chinese characters based on the proposed methods. The experimental results show that above 94.12% of the characters can be segmented successfully.

The proposed approach can be implemented at a fast speed with practical value. The processing time for a typical mail piece image is less than 57 milliseconds on PII-300, which meets the real-time requirements.

Segmentation of handwritten Chinese characters is one of the key technologies for a practical offline handwritten Chinese character recognition system. It can be found that the proposed approach is feasible for segmenting handwritten Chinese characters from the destination addresses on the real mail pieces.

Since proper character segmentation requires prior knowledge of which patterns form a meaningful unit, segmentation itself requires the ability of pattern recognition. But the proposed method is a pure structure-based technique. Further study will focus on utilizing character/subassembly recognition and semantic understanding to improve segmentation performance. The segmentation of the connected handwritten Chinese characters is also a challenging problem that should be handled in the future.

Acknowledgment

The authors would like to thank Shanghai Research Institute of China State Post Bureau for the fund support.

References

1. A. Amin and S. Fischer, "A document skew detection method using the Hough transform," *Patt. Anal. Appl.* **3**, 3 (2000) 243–253.
2. R. G. Casey and E. Lecolinent, "A survey of methods and strategies in character segmentation," *IEEE Trans. Patt. Anal. Mach. Intell.* **18**, 7 (1996) 690–709.
3. O. Chutatape and L. Guo, "A modified Hough transform for line detection and its performance," *Patt. Recogn.* **32** (1999) 181–192.
4. Y. Fu, *Study on Structure and Components of Chinese Characters*, Shanghai Education Press, 1993, pp. 108–169.
5. S. C. Hinds, J. L. Fisher and D. P. D'Amato, "A document skew detection method using run-length encoding and the Hough transform," *Proc. ICPR*, Vol. 1, 1990, pp. 464–468.
6. C. Hong, G. Loudon, Y. Wu and R. Zitserman, "Segmentation and recognition of continuous handwriting Chinese text," *Int. J. Pattern Recognition and Artificial Intelligence* **12**, 2 (1998) 223–232.

7. J. Hu and H. Yan, "A model-based segmentation method for handwritten numeral strings," *Comput. Vis. Imag. Underst.* **70**, 3 (1998) 383–403.
8. J. Illingworth and J. Kittler, "A survey of the Hough transform," *Comput. Vis. Graph. Imag. Process.* **44** (1988) 87–116.
9. Y. Lu, "Machine printed character segmentation — an overview," *Patt. Recogn.* **28**, 1 (1995) 67–80.
10. Y. Lu and M. Shridhar, "Character segmentation in handwritten words — an overview," *Patt. Recogn.* **29**, 1 (1996) 77–96.
11. L. Y. Tseng and R. C. Chen, "Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming," *Patt. Recogn. Lett.* **19** (1998) 963–973.
12. Y. Wu and X. Ding, *Chinese Character Recognition: Theory, Methods, and Implementation*, High Education Press, Beijing, 1992.
13. L. Xu and E. Oja, "Randomized Hough transform (RTH): basic mechanisms, algorithms, and computational complexities," *CVGIP: Imag. Underst.* **57**, 2 (1993) 131–154.
14. X. Zhang, *Chinese Character Recognition Technology*, Tsinghua University Press, Beijing, 1992.
15. K. H Zhang, P. F. Shi and Y. Lu, "Line removal and points interpolation in segmentation of handwritten Chinese characters overlapping written line," *Proc. Int. Symp. Signal Processing and Intelligent System*, 1999, Guangzhou, China, pp. 264–267.
16. C. Zhou, *Machine Recognition of Handwritten Chinese Characters*, Science Publish Press, Beijing, 1997.

Yue Lu is a research fellow at the Department of Computer Science, National University of Singapore. He received his B.E. and M.E. degrees in telecommunications and electronic engineering from Zhejiang University, China in 1990 and 1993, respectively, and his Ph.D. in pattern recognition and intelligence system from Shanghai Jiao Tong University, China in 2000. From 1993 to 2000, he was an engineer at Shanghai Research Institute of China State Post Bureau. Because of his outstanding contribution to the development of the automatic mail sorting system — OVCS, he won the Science and Technology Progress Awards issued by the Ministry of Posts and Telecommunications in 1995, 1997, 1998, respectively. Dr. Lu is a member of the IEEE Computer Society.

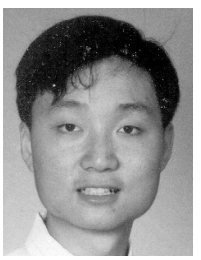
His research interests include document image processing, character recognition, intelligence systems and computer vision.

Pengfei Shi is a Professor and Director at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China. He received his B.S. and M.E. degrees from Shanghai Jiao Tong University in 1962 and 1965, respectively. Prof. Shi is a senior member of the IEEE.

His research interests are image processing, pattern recognition, data mining and artificial intelligence.

Chew Lim Tan is an Associate Professor in computer science at the School of Computing, National University of Singapore. He obtained a B.Sc. (Hons) degree in physics in 1971 from the University of Singapore, an M.Sc. degree in radiation studies in 1973 from the University of Surrey in UK, and a Ph.D. in computer science in 1986 from the University of Virginia, USA.

His research interests are computer vision, document image analysis, intelligent text processing and neural networks.



Kehua Zhang received his B.E. and M.E. degrees from Shanghai Jiao Tong University, China in 1997 and 2000, respectively. He was a visiting researcher at Vriji Universiteit Brussel, Belgium from April–September 2000. Currently he is a Ph.D. candidate in School of Electrical and Computer Engineering, Purdue University, USA.

His research interests include image processing, pattern recognition and wavelet application.