The Design and Implementation of a Chinese Financial Invoice Recognition System

MING Delie , LIU Jian , TIAN Jinwen
State Key Laboratory for Image Processing and Intelligent Control
Institute for Pattern Recognition and Artificial Intelligence
Huazhong University of Science and Technology,
Wuhan, Hubei province, 430074, P.R.China
Tel: 86-27-87556301
Email address: mingdelie@263.net mingdelie@yahoo.com

ABSTRACT

This paper designs and implements a financial invoice recognition system based on the features of the Chinese financial invoice. By using the linear whole block moving method in each vertical segment, a new fast algorithm is put forth to detect and rectify the slant image. To distinguish the different form types (the foundation necessary for locating the form fields, filtering the form lines, etc.), several representative form features are discussed and an invoice-type features library is built by using a semi-automatic machine study method. On the basis of the recognized invoice type, real invoice form is re-oriented against the corresponding blank form according to the invoice type feature, solving the problem of adhesion of characters and form lines, as well as the problem of characters segmentation and recognition. Based on the financial Chinese invoice image feature, a mutual rectification mechanism founded on the recognition results of financial Chinese characters and Arabic numerals is put forward to raise the recognition rate. Finally, the experimental results and conclusions are presented.

Keywords: financial invoice recognition, characters segmentation, optical character recognition (OCR), mutual rectification mechanism, slant-image rectification

1. INTRODUCTION

Invoices, as one of the main information carriers, exist in many aspects of our daily lives. Banks, revenue bureaus, finance and accounting departments use all kinds of invoices as a medium for fund circulation and account settlement. Although these departments have long used computers to process data, billing data are still inputted manually from a keyboard. With the development of the social economy, the number of invoices will grow dramatically. The slow manual data input method has become a bottleneck of high-speed data process.

Besides high speed inputting, the requirement of accurate data entry is essential. Not a single numeral error is acceptable. Any error must be immediately corrected, or the account cannot be settled.

In recent years, with the development of Optical Character Recognition (OCR) technology, the techniques of handwritten Chinese characters and Arabic numerals recognition have made great progress. Ming Delie (2000) reported that the accurate rate of single handwritten financial Chinese characters reached 99%, and the rate of single unconstrained handwritten numerals was above 97%. Single character recognition speed is higher than 100 characters per second on a Pentium II PC. Because Chinese characters and Arabic numerals are filled in columns on invoices, it is very possible to have automatic inspection after recognition.

With the fast development of OCR technology and Document Analysis and Segmentation technology, invoice processing has come to be an important branch of image processing and pattern recognition.

This paper proposes an integrated set of solution blue print on Chinese financial invoice recognition, which can automatically input financial Chinese characters and Arabic numerals from invoices into the computer.

(Figure 1 shows the flow chart of the bill recognition system).

The dot matrix image of an invoice is scanned into the computer. After preprocessing, the whole image is adjusted to horizontal by reversing according to the calculated inclination angle, and the noise in the image is also removed. Then based on the invoice feature extraction, the invoice type recognition is done. If the invoice is an unknown type, the feature data of the invoice is extracted and stored into a Form Type Feature Library. Otherwise, under the guidance of the form type feature, the handwritten financial character fields and numeral fields are separated from the background of the invoice. Individual characters are then separated and recognized. However, the invoice recognition process does not end here. Before inputting the recognition result into computer database, a mutual rectification mechanism based on the one-to-one correspondence relationship between financial Chinese characters and Arabic numerals, is applied to the recognition results to raise the recognition rate.

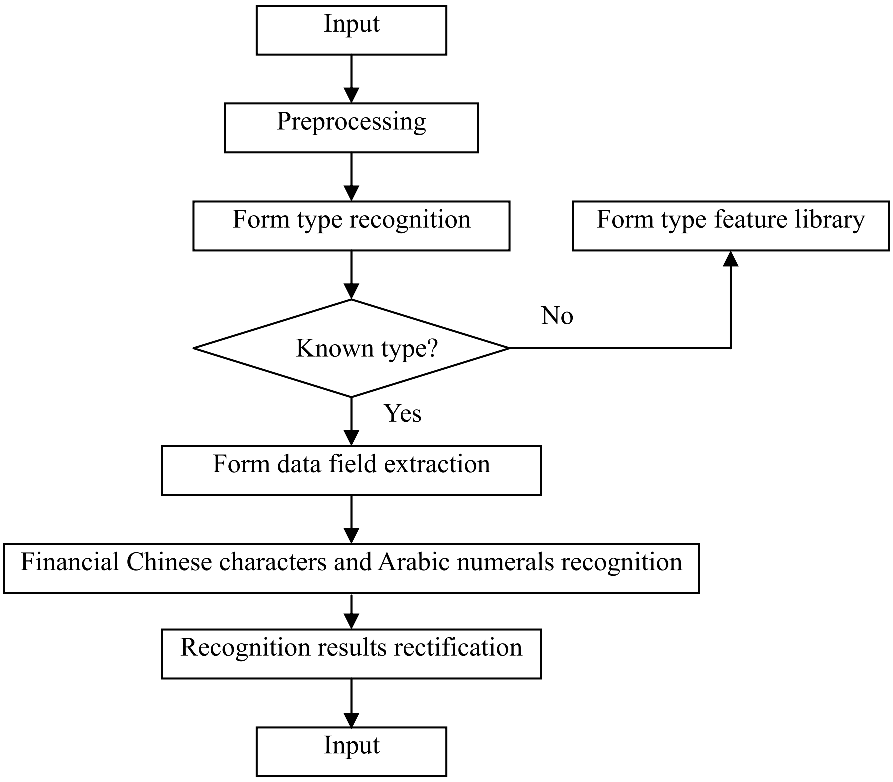

Input
Preprocessing
Form type recognition
Form type feature library
Known type?
No
Yes
Form data field extraction
Financial Chinese characters and Arabic numerals recognition
Recognition results rectification
Input

Fig. 1 Flow chart of bill recognition system

2. PREPROCESSING

A financial invoice usually has items such as bill name, special icon, and form columns. This paper discusses the kind of invoice that can be processed as binary value. The preprocessing includes three parts.

2.1 Smooth processing

As soon as the form image is scanned into the computer, a trembling processing is carried out in the directions of up, down, left, and right. This process can rid disturbance effectively and make the image fringe smoother. Compared with traditional median filter, the trembling process can get the same result in comparatively less time.

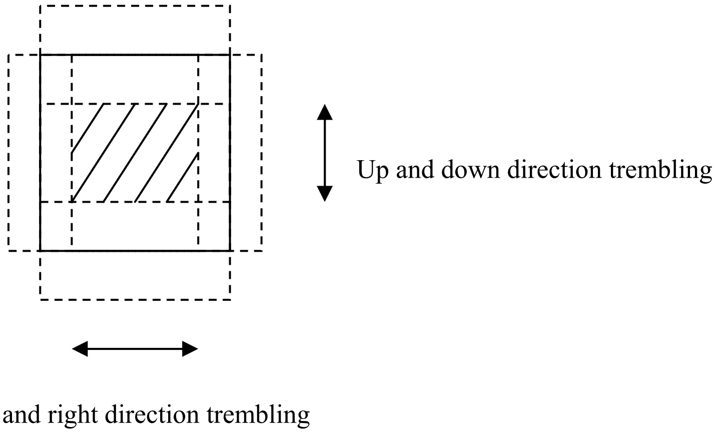

Up and down direction trembling
left and right direction trembling

Fig. 2 Trembling processing sketch map

2.2 Image inclination angle testing

Image slant in the financial invoice recognition system can be divided into two types. One is the whole image slant that is caused by the scanning process and the invoice's printed quality and the other is local handwritten character's slant caused by the personal handwriting quality. We will discuss the two types of slant in the following part.

2.2.1 The whole image's inclination angle testing

Based on the features of the text image in character recognition system, a new algorithm is put forth to detect and rectify the slant image. By using the linear whole block moving method in each vertical segment, this algorithm effectively overcomes the unfavorable fuzzy influence brought by the geometrical rectification, and greatly enhances the executive speed of slant rectification. During the rectification process, this algorithm keeps the forms topology structure unchanged by adding points in the rotating processing thereby facilitating the succeeding location process for each interested form field.
The following is the algorithm for calculating the inclination angle of the image.

A. Scan the image row by row, search for the horizontal line segments using equation 1,

$$(\sum_{i=1}^{n} x_i > \alpha \cdot N) \cdot or \cdot [(\sum_{i=1}^{j} x_i = j) \cdot and \cdot (j \geq m)] \qquad (1)$$

where x_i=1, stands for image information. *N* is the width of image, $n \leqslant N$,0 < $\alpha \leqslant$ 1, 0 < $m \leqslant N$, and the value of n ,α and *m* are obtained by studying the same type bills.

B. Determine succession of the adhesive line segments. If they are in succession, join all contiguous segments into one segment and mark it.

C. Line slope calculation

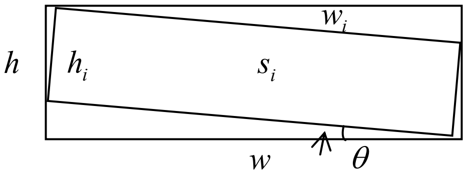


Fig. 3 Line slope calculation model

As shown in Figure 3, the rectangle stands for a line *i*. If the width of the rectangle is w_i, height is h_i ($w_i >> h_i$), slope is k_i, area is s_i, and the number of rows and columns occupied by line *i* are *h* and *w* respectively, then area *s* can be presented as

$$s = w \cdot h = s_0 + k_i \cdot (h_i^2 + w_i^2) \qquad (2)$$

From equation 2, k_i can be obtained easily as

$$k_i = (wh - s_i)/(h_i^2 + w_i^2) \qquad (3)$$

D. Image inclination slope
The image inclination slope is obtained from Equation 4,

$$k = \sum_{i=1}^{n} w_i k_i / \sum_{i=1}^{n} w_i \qquad (4)$$

where *n* is the number of horizontal lines of the whole image.

2.2.2 Local handwritten characters' inclination angle testing

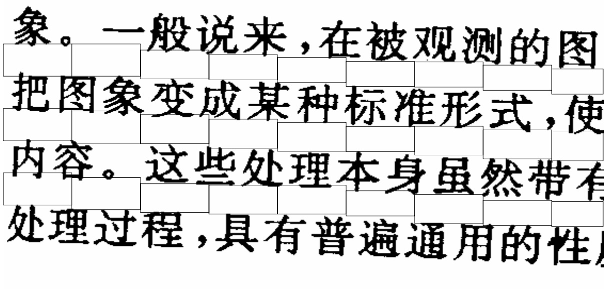
The following is the algorithm for calculating the inclination angle of local handwritten characters. Detailed algorithm implementation can also be found in Ming Delie(2000).

A. Tracing the blank rectangle bar
First, segment the image matrix vertically. The width of each segment depends on the form image quality. To get the accurate inclination angle of the image, the width of each segment must not be set too wide. In our system, we set the width of each segment to 20 pixels (300DPI).
Second, from top to bottom, scanning the image row by row in each vertical segment, searching for the blank rectangle bar

(the blank rectangle bar – Consecutive background area in each vertical segment).

Third, filter some irrespective blank rectangle bars. Because not all of the blank rectangle bars searched out in the last step contribute to the calculation of the inclination angle of the image, we must remove the blank rectangle bar that may disturb following the calculation and increase the unnecessary quantum of calculation.

Figure 4 shows the blank rectangle bar tracing result in a form field.



象。一般说来，在被观测的图
把图象变成某种标准形式，使
内容。这些处理本身虽然带有
处理过程，具有普遍通用的性

Fig. 4 The blank rectangle bar tracing result

B. Calculating the average image inclination fall

To illustrate this step, inspect the procedure of calculating the average image inclination fall between M $_1$ and M $_2$ which are adjacent vertical segments(M $_2$ = M $_1$+1). If n $_1$ and n $_2$ are the respective bar number of M $_1$ and M $_2$, using the little one as the radix(assuming n $_1$ < n $_2$), to each blank bar in M $_1$, searching for the corresponding adjacent blank bar in M $_2$. The rule of correspondence lies in the following two points. First, the heights of the two blank bars are similar. Second, the two blank bars are adjacent and interlaced vertically.

After matching each bar between M $_1$ and M $_2$, the fall Δh_{12} between the two segments can be calculated out by formula 5.

$$\Delta h_{12} = \sum_{i=1}^{n} (x_i - y_i) \Big/ n \qquad (5)$$

In formula 5, x_i and y_i are the central lines of a pair of matched bar between two adjacent segments, n is the total pair number of matched bar.

The whole image inclination fall is presented in formula 6. In it, n is the number of vertical segments.

$$S = \sum_{i=1}^{n-1} \Delta h_{i,(i+1)} \qquad (6)$$

Through S calculated by formula 6, we can judge whether the image is inclined and calculate the exact inclination angle by the following rule.

$$\begin{cases} S > threshold \text{ — clockwise inclining} \\ -threshold \le S \le threshold \text{ — no inclining} \\ S < -threshold \text{ — anti - clockwise inclining} \end{cases}$$

2.3 Image slope adjustment

The image slope adjustment is actually a coordinate rotate transformation. Assume that *(x, y)* is the source coordinates of point *P*, and *(x', y')* is the object coordinates, then the image slope adjustment of random slant angle of θ is processed by using formula 7.

$$\begin{cases} x' = x \cdot \cos\theta - y \cdot \sin\theta \\ y' = y \cdot \cos\theta + x \cdot \sin\theta \end{cases} \qquad (7)$$

In the image the positive direction of y-axis is downwards, so the positive direction of θ is clockwise.

We have developed a fast rotate algorithm for images with slope less than 5 degrees. The invoice image's topology structure is kept unchanged by adding points in the rotating process.

Now we assume the invoice image inclines clockwise, and the fall between the two side of the image is S. We cut the image to S segments vertically. The width of each segment is Image-Width/S. From left to right, we lift each segment up K pixels(K=1,2,….,S). In order to keep the form's topology structure unchanged, we add points in the lifting processing. After rectifying the image vertically, we must do the same horizontally with the top-to-bottom fall S_1 (=Image-Height*S/ Image-Width). Figure 5 is the sketch map of the linear whole block moving algorithm.

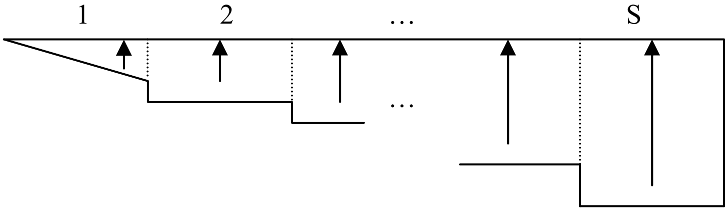


Fig. 5 The linear whole block moving algorithm

3. FORM TYPE FEATURE EXTRACTION AND RECOGNITION

E.Green(1995) mentioned that, in order to put the system into practical use, it must have the ability to distinguish different bill types, locating the form fields, filtering the form lines, etc. To achieve this target, a form feature library is established to record the features of any new type of form entered into the system. In the process of the construction of the form feature library, correct form feature extraction is the key.

In the process of form feature extraction, an invoice form study mechanism is set up to let the computer remember all types of form type features. The form study serves three functions. First, it can recognize the invoice image inputted according to type features. Second, it can locate the filled fields of an invoice form at a high speed extracting the useful information from the filled fields. And third, it can reconstruct the invoice form with high quality.

3.1 Extracting primary features of bill image interactively

A well-printed blank invoice is used as the study sampler. After scanning into computer, the invoice will go through smooth and slope adjustment pre-processing. The invoice's name field and special icon field are specified interactively. The computer establishes the original point of image, and marks the name field, special icon field and the coordinate positions of the out frame of the form. The computer automatically searches for the horizontal and vertical lines from the form image. After the computer finishes the study process, the study results are displayed and will be confirmed manually item by item.

3.2 Identifying fixed field features

The fixed field features are the printed sections of an invoice form. The sections are designed for the invoice managers to get the needed information. They usually include the form name, special icon, form columns, description of the contents in each column, verification rules and etc. Other commonly used features include numbers of vertical and horizontal lines, the space between adjacent form lines, the size and location of special icon field, the number of columns and other original printed information on the invoice images.

Besides the features mentioned above, a new type of form feature is put forward to reflect the planar distribution of form texture which is made up of form lines, special icons, form titles and other originally printed form information. In practical use, this new feature has been proven to be stable and correct, and most importantly, can be easily extracted.

The purpose of identifying the fixed field features is to describe a blank invoice form (an unused form) exactly by using a suitable data structure, and to create blank form information file. The form type features library will be stored in a file as the basis for quantitative analysis in the following process for each item in the invoice.

Byte Number	Content
1 ~ 2	Scan precision (dpi/mm)
3 ~ 10	Name field
11 ~ 18	Special icon field
19 ~ 20	Number of vertical lines
21 ~ 22	Number of horizontal lines
23 ~ 24	Number of columns
25 ~ 26	Length of record2 file

Table1: Blank form identification information

We have designed two record files to store the information of the fixed features. File record1 is a file of fixed size, while record2 is a file of variable size. The size of record2 is stored in bytes 25-26 of record1. The file format of record1 is shown in table 1. The size of the name field and the special icon field are expressed by the top-left corner coordinate and bottom-right corner coordinate of its frame. If there is no special icon field, the corresponding column of Table 1 is filled by zero. Record2 is created according to the following algorithm:

A. Record the coordinates of every vertical or horizontal line's top left corner and bottom right corner in turn. For the horizontal line, Δx is its length and Δy is its width, while the vertical line, Δy is its length and Δx is its width. Finally store the number of vertical lines and horizontal lines in bytes 19-20 and bytes 21-22 of record1.

B. Using columns traversal algorithm to code the columns one by one. Record the coordinates of every column's top left and bottom right corners.

C. Search for every column and query the number of characters in every column. Store the number in the column attribute table. Record the coordinates of the top left and bottom right corners of circumventing rectangle of every character.

D. Extract the distance between the edge of the name field and the special icon field and the circumventing rectangles of these fields. This distance will be used in the invoice type recognition.

Figure 6 is a sample blank form, which is created after machine study.

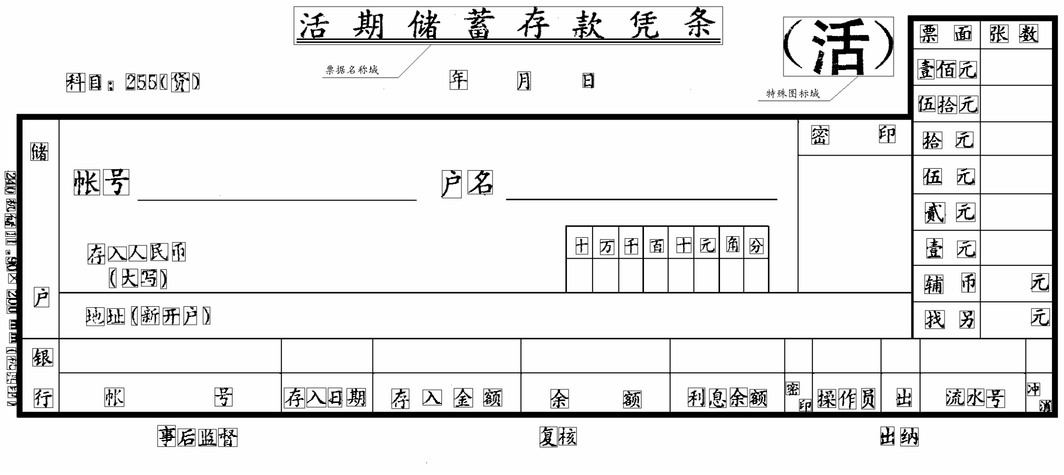


Fig. 6 A sample blank form

4. LOCATING REAL FORM AGAINST BLANK FORM

Real form is an invoice form with data filled in according to the bill identifier column. The differential form is the remainder of a real form after the blank form section has been filtered out. The differential form will be recognized by computer.

4.1 Real form type recognition

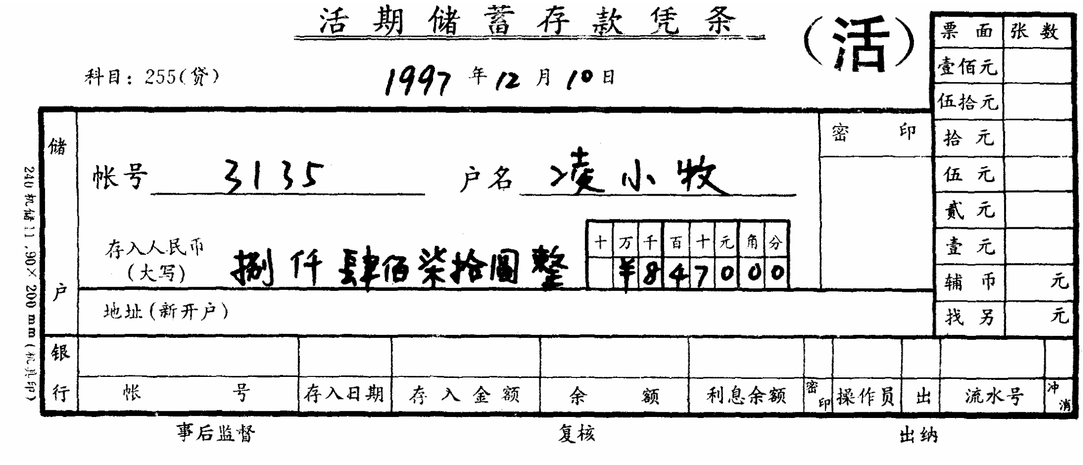
Load the type features of all invoice form types into the computer, matching them with the newly imported real form by their name field, special icon field and circumventing frame size. The matching weight coefficient is higher for the special icon field feature. Use the form type with minimum distance as the result for the type of invoice to be recognized.

4.2 Re-orienting real form against blank form

The real form is interpolated according to the recognized form type and its type features. In the interpolation process, be attentive to eliminating the cumulative errors and adding points to the relevant portion to ensure that the form lines and the topology structure of characters to be recognized are unchanged. After the real form is re-oriented against the blank form, the invoice name field and special icon field are eliminated first. The thickness of the blank form's horizontal lines and vertical lines are expanded to one circle and the corresponding background field in the real form is filtered out. Finally all characters in the identified column are eliminated. When eliminating the form's horizontal and vertical lines, take special care of the adhesion between character and form line segment.

If a horizontal (or vertical) line crosses the strokes of the characters, they can be correctly easily by combining the up and down strokes if they are cut by a horizontal line, or by combining the left and right strokes if they are cut by a vertical line. The processing will become more difficult if the form lines are tangential to character strokes. In this case we should examine how much the line's skeleton deviate from horizontal (or vertical) line's center, and then combine the tangential stroke segments by adding points accordingly.

After the above processing, we can obtain a differential form that is only constructed by user filled parts. Figure 7 is an example of the real form relevant to the type that is shown in Figure 6. Figure 8 is a differential form that is obtained by eliminating the blank form information from the re-oriented real form.



活期储蓄存款凭条 （活）

科目：255（贷） 1997 年 12 月 10 日

240机储11.90×200mm（机具印）

储户

帐号 3135 户名 凌小牧

存入人民币（大写） 捌仟肆佰柒拾圆整

十	万	千	百	十	元	角	分
	¥	8	4	7	0	0	0

密 印

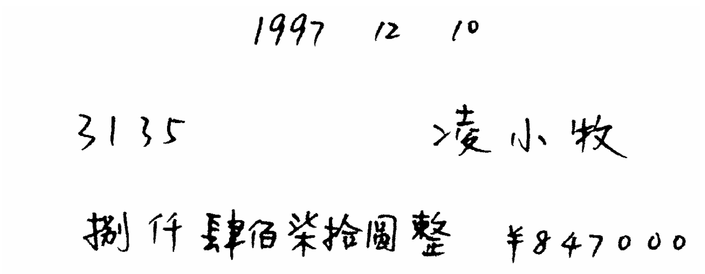
地址（新开户）

票面	张数
壹佰元	
伍拾元	
拾元	
伍元	
贰元	
壹元	
辅币	元
找另	元

银行									
行	帐号	存入日期	存入金额	余额	利息余额	密印	操作员	出	流水号	冲消

事后监督 复核 出纳

Fig. 7 An example of the real form relevant to the type shown in Figure 4



1997 12 10

3135 凌小牧

捌仟肆佰柒拾圆整 ¥847000

Fig.8 A differential form that is obtained by eliminating the blank form information from the re-oriented real form

5. CHARACTERS SEGMENTATION AND RECOGNITION

In section 4, our aim is to accurately locate each form field to extract an intact character string in each field. Based on the recognized form type, the real form whose fields are filled with information character strings are re-oriented against the corresponding blank form. This will contribute to solving the problem of adhesion of characters and form lines, and the form lines and other intrinsic form information can also be easily filtered. The following step segments the character string extracted from each form field and passes each individual character to the OCR engine.

Because the method of single Chinese character and numeral recognition has been discussed by Lou Xiaoping(1998) and Hu Jiazhong(1996), it won't be described here in detail. It is remarkable, however, because several methods, such as background features, the distribution of outline's convex and concave, local field features and left right profile different features were used in the numeral recognition. Thus, we were able to remark the reliability of the recognition result.

There may be adhesion between neighboring handwritten financial Chinese characters. On the other hand, some financial Chinese characters are constructed by left and right parts (e.g. Chinese equivalent of the Arabic numeral “ten”), or even by left, middle and right parts (e.g. Chinese equivalent of the Arabic numeral “eight”). In these cases, to correctly segment every financial Chinese character becomes critical in order to enhance the accurate recognition rate of the system. In this system, two measurements are adopted to ensure there is no segmentation error.

A. According to the length of the Arabic numeral string, search for the relevant unit words within financial Chinese characters. Use the minimum distance principle to separate these unit words.

B. When recognizing single financial Chinese character, raise the weighted coefficient of distance of every financial Chinese character’s left and right strokes. Take the minimum value of total distance of the whole character string as a standard, then determine the final recognition result.

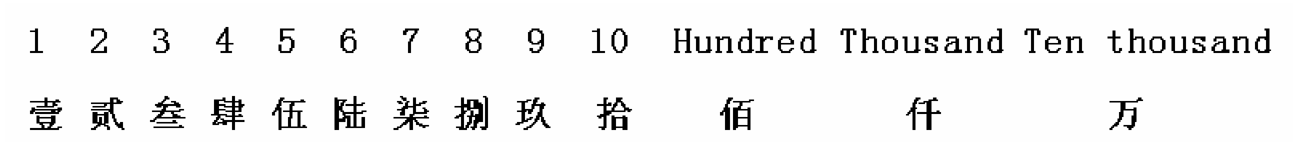
After the financial characters and Arabic numerals are recognized separately, we can determine the final recognition result (the sum of a typical invoice image) according to the following two factors:

A. The reliability of every financial Chinese character and the relevant Arabic numeral.

The reliability measurement of each financial Chinese character and the relevant Arabic numeral as presented by the recognition algorithm.

B. The corresponding relationship between Chinese character and Arabic numeral presentation of the sum of the bill image.

The corresponding relationship between financial Chinese characters and Arabic numeral in commonly used financial Chinese invoices is shown in figure 9.



1	2	3	4	5	6	7	8	9	10	Hundred	Thousand	Ten thousand
壹	贰	叁	肆	伍	陆	柒	捌	玖	拾	佰	仟	万

Fig.9 The one-to-one corresponding relationship between financial Chinese characters and Arabic numeral in financial Chinese invoices

Based on the financial Chinese invoice image feature, a mutual rectification mechanism based on the recognition results of financial Chinese characters and Arabic numerals is put forward to raise the recognition rate. Figure 10 is an example of the real form relevant to the bill that is shown in Figure 7.

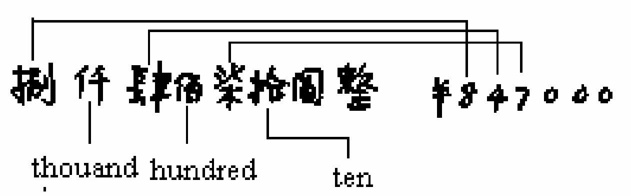


Fig.10 One-to-one corresponding relationship sample in real financial bills

After the process of the mutual rectification mechanism, the final recognition result is inputted into computer database and participates in all kinds of statistical operations.

6. EXPERIMENTAL RESULTS AND CONCLUSIONS

In detailed technical approach, the system has been programmed and implemented according to the aforesaid proposal. We have implemented and tested the system with more than 1000 pieces of form in more than ten different types used in different departments and gained a satisfying result. Table 2 shows the experimental results.

Recognition Results’ Accuracy Rate Invoice Type	Accuracy rate of single financial handwritten Chinese character (%)	Accuracy rate of single financial handwritten Arabic numeral (%)	Accuracy rate of the ultimate invoice’s par value before rectification (%)	Accuracy rate of the ultimate invoice’s par value after rectification (%)
Type 1	96.62	93.19	97.42	99.20
Type 2	96.95	92.80	97.00	99.79
Type 3	97.20	92.87	97.33	99.23

Type 4	96.50	94.34	98.12	100.0
Type 5	96.37	92.00	98.74	99.48
Type 6	97.10	90.19	97.14	99.75
Type 7	96.78	95.10	97.98	99.37
Type 8	96.00	92.74	97.10	99.26
Type 9	95.61	96.01	98.17	100.0
Type 10	93.12	95.20	96.88	99.04

Table2: Experimental test results

The test results shows that after adopting the supervised studying method to acquire the form type features library, the image of the blank form can be eliminated from a real form effectively. The accuracy rate of single financial Chinese character and Arabic numeral recognition keeps steady. After the mutual rectification process between the recognition result of financial Chinese characters and Arabic numerals, we can raise the invoice's recognition accuracy rate significantly. The adhesion of form line and characters make character segmentation difficult, becoming a major factor in increasing the recognition rate. However, further improvement is still needed.

REFERENCE

Ming Delie, Liu Jian, Hu Jiazhong, 2000. An Improved Algorithm for Handwritten Chinese Text Segmentation. Journal of Huazhong University of Science and Technology 28(2), pp.87-89.

Ming Delie, Liu Jian, Hu Jiazhong, 2000. The Slanting Image Detecting and Rectifying Technology in OCR System. Journal of Huazhong University of Science and Technology 28(5), pp.66-68.

E.Green, M.Krishnamoorthy, 1995. Model-Based Analysis of Printed Tables. In: International conference on Document Analysis and Recognition (1), pp. 214-217.

Lou Xiaoping, Hu Jiazhong, 1998. The Segmentation and Recognition of Unconstrained Handwritten Numerals. In: The Latest Development of Computer Intelligent Interface and Intelligent Applications (1), pp. 86-90.

Hu Jiazhong, Yang Xiaofei, 1996. Handwritten Chinese Character Recognition System. Journal of Software (10), pp. 15-20.

S.Mori and T.Sakura, "Line Filtering and its Application to Stroke Segmentation of Hand-printed Chinese Characters", Proc. of the 7th Intl. Conf. on Pattern Recognition, pp.366-369, 1984.

Pavlidis T., "A Vectorizer and Feature Extractor for Document Recognition", Computer vision, Graphics, and Image Processing, No.35, pp.111-127, 1986.

H.Bunke, "Automatic Interpretation of Text and Graphics in Circuit Diagrams", Pattern Recognition Theory Applications,pp.297-310,1982.

M.Karima, K.S.Sadahl, and T.O.McNeil, "From Paper Drawings to Computer Aided Design", IEEE Computer Graphics and Applications, pp.24-39,Feb.1985.

L.A.Fletcher and R.Katsuri, "Segmentation of Binary Images into Text Strings and Graphics", SPIE Vol.786 Applications of Artificial Intelligence, pp.533-540,1987.

C.C.Shih and R.Katsuri, "Generation of a Line Description File for Graphics Recognition", SPIE Vol.937 Applications of Artificial Intelligence, pp.568-575,1988.

A.Pizano, "A Line Recognition Algorithm for Business Form Processing", Technical Report SRC901101, Ricoh Corporation Software Research Center, Santa Clara, CA, Nov.1990.