Acta Technica *63* No. 1B/2018, 1–8

Invoice number recognition algorithm based on digital structure features

Junjie Liu[1], Jianhui Lin[1,2]

Abstract. A new recognition algorithm based on digital structure features is proposed to solve problems of poor anti-interference ability in the template matching algorithm and long time-consuming of neural network algorithm. The Multi-Scale Retinex algorithm is used to eliminate the uneven brightness of invoice images, and the noise is removed by adaptive medium filter. Aiming at the problem of digital adhesion, the area threshold method is proposed to segment accurately. After the normalization of numbers, topology features are introduced to complete the fast classification of numbers. Compared with the traditional recognition algorithm based on digital structure features, the new algorithm choose more typical digital features and reduce the number of extracted digital features. Experimental results show that this algorithm has faster recognition speed and higher recognition rate.

Key words. Digital structure features, uneven brightness, image segmentation, area threshold method, topology features.

1. Introduction

As a business voucher to record business activities, invoices are the important evidence for enterprises to protect their rights and interests according to law. By making use of fake invoices, criminals have caused huge losses to the national finance and enterprises every year. Every value-added tax invoice has unique number information. The number information of each Chinese value-added tax invoice is made up of 18 digits, with 10-digit invoice code and 8-digit invoice number. The enterprise usually records the invoice number information manually and sends the invoice number information to the national tax website to check the authenticity of the invoice. The computer automatically identifies the invoice number by scanning or photographing and uploads it to the website to check its authenticity. As a part of OCR technology [1–2], digital recognition technology has high research values, recognition rate and speed are used as evaluation criteria for digital recognition al-

[1]School of Technology, Beijing Forestry University, Beijing, 100083, China
[2]Corresponding author

gorithms. With the development of digital recognition technology, many algorithms have been proposed, such as recognition method based on template matching proposed by [3], neural works recognition method proposed by [4], recognition method based on structural features proposed by [5] and other algorithms. Template matching algorithm sets the standard template and calculates the Euclid Distance between templates and objects to classify. The neural network algorithm has the characteristics of self-adaptation and learning ability, it can guarantee high accuracy for the broken number, but the amount of calculation is large and the sample training time cannot meet the real-time requirements. The traditional recognition algorithms based on structural features calculates the Euclid Distance between digital structure features and templates to classify. This algorithm reduces the computation, but the number of extracted digital features is too much, the digits which are seriously damaged cannot be identified accurately. Therefore, this paper proposes a new algorithm based on digital structure features, which introduces topological measure to reduce the number of extracted features, and uses less digital features to complete classification. Compared with the traditional algorithm based on digital structure features, the new algorithm can reduce the computation and improve the accuracy rate.

2. Preprocessing

This paper focuses on the identification of invoice numbers. Because of the unavoidable contamination in the image captured by cameras or mobile phones, preprocessing is necessary and the specific flowchart is shown in Fig. 1.

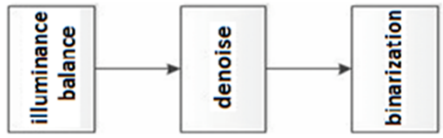

Fig. 1. Flowchart of preprocessing

Flowchart of preprocessing

2.1. Illuminance balance

The image of invoice captured by the camera will inevitably be affected by the uneven illumination of the environment. The Multi-Scale Retinex algorithm is applied to solve the problem of uneven illumination. The Retinex algorithm proposed by Jobson et al. in 1997 solved the problem of uneven brightness [6]. The image S saw by the naked eyes is composed of two parts: the reflection component R which carry the detailed information and the incident component L of the ambient light which is the interference factor. Their relationship is shown as follows:

$$S(x,y) = R(x,y) \times L(x,y) \tag{1}$$

$$\mathrm{Log}\left[R\left(x,y\right)\right] = \mathrm{Log}\left[S\left(x,y\right)\right] - \mathrm{Log}\left[L\left(x,y\right)\right] \tag{2}$$

The component R is finally obtained by estimating the component L and removing the component L according to the image S. The component R represents the real appearance of objects. In Multi-Scale Retinex algorithm, the incident component is estimated and eliminated by Gauss filtering.

The task consists of the following steps:

1. Select three fuzzy radiuses: 15, 80, 200.

2. Calculate the image $S(x, y)$ filtered by three fuzzy radiuses respectively.

3. Calculate $\mathrm{Log}[R(x, y)]$ by formula (2).

4. Quantizing the pixel values of $\mathrm{Log}[R(x, y)]$ to the range of 0 to 255.

2.2. Denoise

The images captured by cameras must be affected by noise. The noise of image is mainly distributed in the high frequency region; in the meantime the edges and details of the image are also distributed in the high frequency region. Because adaptive median filters can remove noise effectively and retain details of image, the adaptive median filter proposed by [7] is used to remove the noise in this paper.

2.3. Binarization

Otsu algorithm is used to obtain binary image. The steps of Otsu algorithm are showed as follows:

(1) Calculate the normalized histogram of image and use $p_j,\ j = 0, 1, 2, \ldots, L-1$ to represent the components of the histogram.

(2) Calculate cumulative sum $q(k),\ k = 0, 1, 2, \ldots, L-1$ by formula

$$q\left(k\right) = \sum_{i=0}^{k} p_i\,. \tag{3}$$

(3) Calculate cumulative mean value $m(k),\ k = 0, 1, 2, \ldots, L-1$ by formula

$$m\left(k\right) = \sum_{i=0}^{k} ip_i\,. \tag{4}$$

(4) Calculate global gray mean m_{G} by formula

$$m_{\mathrm{G}} = \sum_{i=0}^{L-1} ip_i\,. \tag{5}$$

(5) Calculate the variance $\sigma_{\mathrm{B}}^2(k),\ k=0,1,2,\ldots,L-1$ by formula

$$\sigma_{\mathrm{B}}^2(k)=\frac{[m_{\mathrm{G}}q(k)-m(k)]^2}{q(k)[1-q(k)]}. \tag{6}$$

(6) The threshold value τ is the k value that makes the maximum variance: $\sigma_B^2(\tau)=\max\left(\sigma_B^2(k)\right),\ k=0,1,2,\ldots,L-1$. If the maximum value is not unique, the average value of each k is selected as the threshold value. The binary image is shown in Fig. 2.

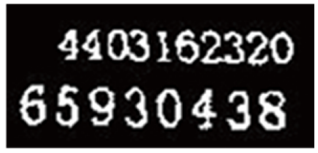

Fig. 2. The binary image

3. Digital recognition

Firstly, each digital image is normalized to size of 24×42, and then numbers are identified according to the structural features of the numbers. The logic of digital classification is shown in Fig. 3

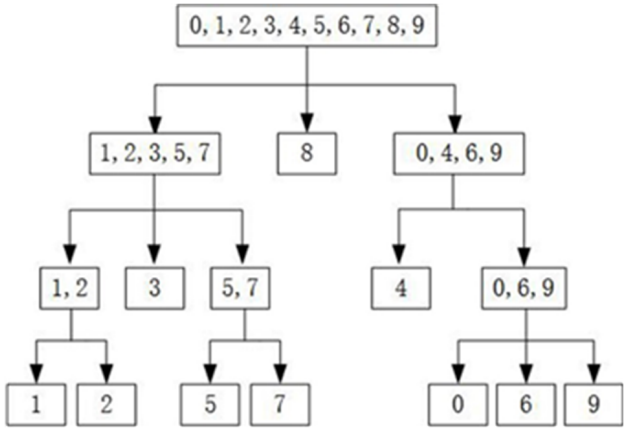

Fig. 3. The classification tree

According to the above figure, the task consists of the following steps:

(1) After observation and analysis of the both forms of figures topology structure, using the Euler number characteristics as the first level classification, the numbers are divided into 3 groups: (1, 2, 3, 5, 7), (8), (0,4,6,9) and the value of Euler number is 1, -1, 0, respectively.

(2) We continue to classify the group (1, 2, 3, 5, 7) into smaller groups based on characteristics of upper lines and lower lines. The characteristics of upper lines and

lower lines are shown in Fig. 4

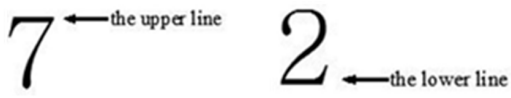

Fig. 4. Characteristics of upper lines and lower lines

(3) Analyzing the vertical projections of numbers to determine whether there is vertical line. The vertical projections of the number 1, 2, 5, 7 are shown in Fig. 5.

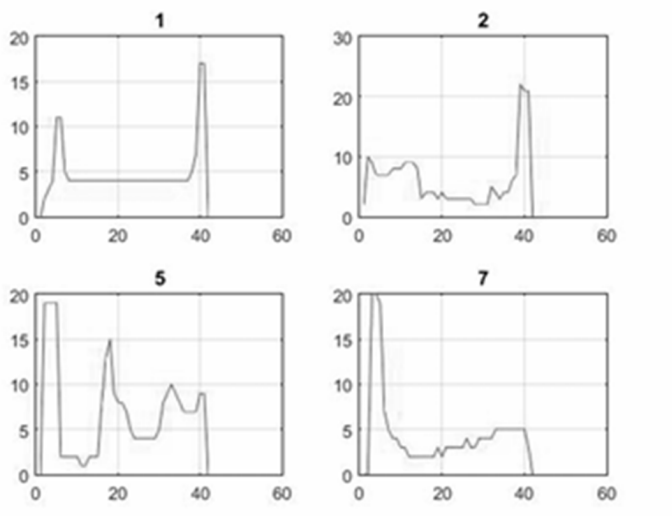

Fig. 5. Horizontal projection of numbers

For the normalized 24×42 size digital images, calculate the value of V_n as

$$V_n = L_n \text{div } H, \quad (n \in [1, 42]), \tag{7}$$

where $H = 24$ is the total length of a column in horizontal projection and n is the number of columns in horizontal projection and L_n is the sum of pixels in the nth column. If $V_n \in [0.6, 1]$, we think the number has a horizontal line. Define 1st to 5th column of horizontal projection as the top region of number and define 38th to 42th column of horizontal projection as the bottom region of number. If there exists $V_n \in [0.6, 1]$ when n ranges from 1 to 5, it is thought that the number has an upper line. On the contrary, if there exists $V_n \in [0.6, 1]$ when n from 38 to 42, it is thought that the number has a lower line.

Thus the (5, 7) is classified into a group based on the upper line, (1, 2) is divided into a group based on the lower line. The number 3 is divided into a group without an upper line or a lower line.

(4) In order to continue the downward classification, we set the characteristics of the vertical line as the following figure shows Fig. 6

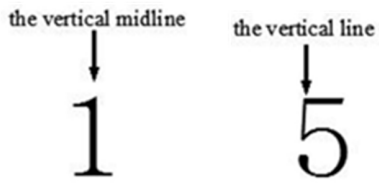

Fig. 6. Characteristics of vertical lines

Analyzing the vertical projections of numbers to determine whether there is vertical line. The vertical projections of the number 1, 2, 5, 7 are shown in Fig. 7.

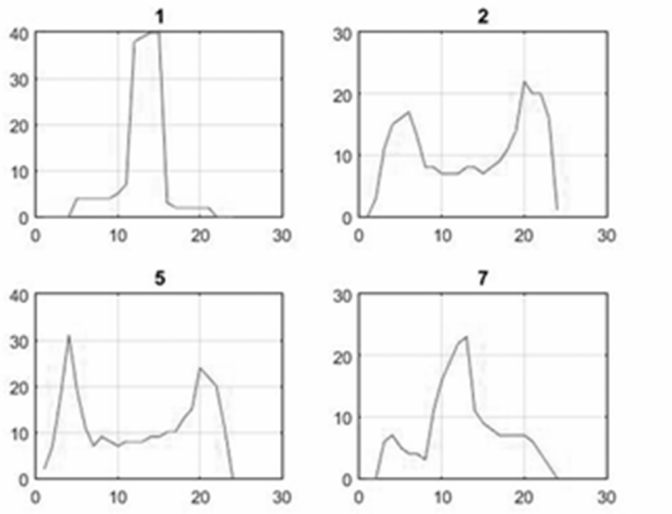

Fig. 7. Vertical projection of numbers

In the group of (1, 2), using the characteristics of vertical midline to distinguish 1 from 2. For the normalized 24×42 size digital images, calculate the value of T_n

$$T_n = h_n \text{div } w \quad (n \in [1, 24]), \tag{8}$$

where $w = 42$ is the total length of a column in vertical projection and n is the number of columns in vertical projection and h_n is the sum of pixels in the nth column. If there exists $T_n \in [0.7, 1]$ when n ranges from 10 to 20, it is thought that there is a vertical midline and the number is 1, otherwise it is the number 2.

In the group of (5, 7), use the characteristics of vertical lines to distinguish 5 from 7. If there exists $T_n \in [0.6, 1]$ when n ranges from 0 to 10, we believe that there is a vertical line, it is the number 5. Otherwise it is the number 7. The numbers are divided into two parts from the middle of the image in vertical direction which are shown in Fig. 8

By observing the area distribution of connected domains, the area of number 9 is mainly concentrated in the right part of the figure, while the area of the number

Fig. 8. Images of digital after dividing

6 is mainly concentrated in the left part of the figure and the areas of two parts of number 0 is almost equal. The area difference between the left half and the right half is calculated and the threshold of area is set to 30, if the difference is greater than the threshold value, it is considered as 6 or 9. When the area of the right part is larger, it is considered as the number 9. Otherwise, it is the number 6. If the difference is less than the threshold value is considered as the number 0.

4. Experimental results

Ensure software version, running device and character objects are consistent and use the same preprocessing. Recognition speed and accuracy are utilized as evaluation criteria. Experimental software adopts the Matlab R2015a, Window 7 operating system, i5-2520M processor. 900 numbers from invoices are recognized by the proposed method, the template matching method [3] and the traditional template feature matching method [5]. The experimental results are shown in Table 1.

Table 1. Experimental results

Recognition algorithm	Template matching	Template feature matching	Proposed algorithm
Number of digits	900	900	900
Accuracy rate	85.55 %	92.88 %	94.88 %
Time (s)	4.3	3.7	3.2

The experimental results show that the algorithm in this paper guarantees the faster speed and high accuracy rate. In this paper, the classification algorithm based on the features of digital structure choose more stable digital structure features and reduce the number of extracted digital features. Therefore, this algorithm has better accuracy and speed.

5. Conclusion

Aiming at the recognition algorithm based on digital structure in this paper, some conclusions can be drawn as follows:

(1) The algorithm in this paper can achieve high recognition efficiency. The recognition rate of numbers on the invoice is 94.88 %, which is higher than the recognition rate of the traditional digital recognition algorithm. This method makes

up for the shortcomings of complex features extraction, it selects the digital features with higher discrimination and more stable structures to classify. Therefore, the accuracy is higher and the speed of recognition is faster.

(2) The Multi-Scale Retinex algorithm is applied to solve the problem of uneven brightness of images. Gauss filer of three radiuses is used to estimate the component of environmental illumination, it strengthens the details of the image and eliminates the influence of uneven brightness at the same time.

(3) In order to solve the problem of invoice digital adhesion, an area threshold method is proposed according to characters of invoice numbers, which has high accuracy for the segmentation of digital adhesions.

References

[1] A. R. Alexandria, P. C. Cortez, J. H. S. Felix, A. M. Giräo, J. B. B. Frota, J. A. Bessa: *An OCR system for numerals applied to energy meters.* IEEE Latin America Transactions *12* (2014), No. 6, 957–964.

[2] R. E. Drinkwater, R. W. N. Cubey, E. M. Haston: *The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels.* PhytoKeys (2014), No. 38, 15–30.

[3] M. Salehifar, T. Nanjundaswamy, K. Rose: *Joint design of layered coding quantizers to extract and exploit common information.* Data Compression Conference (DCC), 30 March–1 April 2016, Snowbird, UT, USA, IEEE Conferences (2016), 631–631.

[4] Y. J. Tang, X. J. Shen, W. I. Zhu, C. H. Sui: *Recognition system for character of numeral instrument dynamic displayed based on BP neural network.* Electrical Measurement & Instrumentation *42* (2005), No. 9, 42–45.

[5] M. Salehifar, E. Akyol, K. Viswanatha, K. Rose: *On optimal coding of hidden Markov sources.* Data Compression Conference (DCC), 26–28 March 2014, Snowbird, UT, USA, IEEE Conferences (2014), 233–242.

[6] D. J. Jobson, Z. Rahman, G. A. Woodell: *Properties and performance of a center/surround retinex.* IEEE Transactions on Image Processing *6* (1997), No. 3, 451–462.

[7] H. Hwang, R. A. Haddad: *Adaptive median filters: New algorithms and results.* IEEE Transactions on Image Processing *4* (1995), No. 4, 499–502.

Received October 12, 2017