A New Method on the Segmentation and Recognition of Chinese Characters for Automatic Chinese Seal Imprint Retrieval

Chao Ren
Graduate School at Shenzhen,
Tsinghua University
Shenzhen, China
renc09@mails.tsinghua.edu.cn

Dong Liu
Graduate School at Shenzhen,
Tsinghua University
Shenzhen, China
liu-d05@mails.tsinghua.edu.cn

Youbin Chen
Graduate School at Shenzhen,
Tsinghua University
Shenzhen, China
chenyb@sz.tsinghua.edu.cn

***Abstract*—In this paper, we propose a new method on the segmentation and recognition of Chinese characters in both circular and elliptical seals for automatic Chinese seal imprint retrieval. The seal identification system compares an input seal image with its model seal. However, the model seal is usually found manually, which costs quite a lot of time and this manual stage has become the bottleneck of automatic seal identification. In addition, traditional automatic methods which were designed for circular seal imprint retrieval cannot work well for elliptical seal images. In order to overcome these shortcomings, our method first fits the contour of seal images and then automatically classifies them into either circular seal shape or elliptical seal shape. After that, we apply different methods for Chinese character segmentation and skew correction with circular and elliptical seal shape images respectively. Finally, we use our classifier combination strategy to recognize the Chinese characters on Chinese seal images. Once optical character recognition (OCR) is done, a model seal imprint corresponding to its input sample seal image is retrieved for verification from the seal imprint database. Experimental results show that our method has better performance than traditional ones.**

Keywords- seal image; ellipse fitting; character segmentation; skew correction; OCR; classifier combination

I. INTRODUCTION

In many eastern countries, a large number of seal images need to be identified every day. Generally an automatic batch mode seal identification system includes several modules, such as seal imprint detection and extraction [1,2,3], seal character recognition [4], seal retrieval [5], and seal verification [6,7]. In such a system when a sample seal image is input, it should be compared with its model seal image which is called as the verification procedure. The model seal is usually found manually and this manual stage has become the bottleneck of automatic batch mode seal identification. Liu et al. [5] introduced an automatic seal image retrieval method by using shape features of Chinese characters. However, their algorithm cannot work well for elliptical seal imprints which are used quite often.

In this paper, we propose a new method to segment and recognize Chinese characters on both circular and elliptical seal images. Our method first fits the contour of seal images and then classifies them into either circular seals or elliptical ones. After that, we apply different methods for Chinese character segmentation and skew correction with circular and elliptical seal images respectively. Finally, we use classifier combination for the OCR of Chinese seal images. Experimental results show the effectiveness of our proposed method.

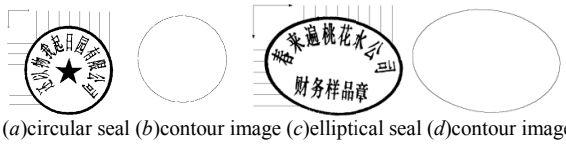
(*a*) circular official seal (*b*)circular specific-use seal (*c*)elliptical seal

Figure 1. Samples of Chinese seal images

Fig.1 shows three different styles of seal images which are often used in China according to the provision [8] of the State Council. Fig.1 (*a*) shows the style of official circular seals standing for the highest duty and power of an entity in which the main characters surround the pentacle in the center. Fig.1 (*b*) and (*c*) exhibit the types of specific-use circular and elliptical seals which are widely used by finance or accounting department of an enterprise. Different from official circular seals, the specific-use circular seals have some characters in a line instead of surrounding the pentacle. These characters forming a character line are considered as the auxiliary region in a seal and they will not be recognized in our proposed method since they are not helpful for the model seal retrieval.

This paper is organized as follows: Section Ⅱ introduces how to automatically distinguish between circular and elliptical seal imprints. Section Ⅲ introduces character segmentation and skew correction for both circular and elliptical seal imprints. Section IV presents the OCR engine combination for the recognition of Chinese seal characters. Section Ⅴ shows the experimental results. Section Ⅵ draws the conclusion.

II. DISTINGUSHING SEAL SHAPES



(*a*)circular seal (*b*)contour image (*c*)elliptical seal (*d*)contour image

Figure 2. Detecting the outer contour of some sample seals

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.197

The extracted binary seal images are shown in Fig.2 (*a*) and (*c*). For all binary images in this paper, we define pixels in black as foreground and pixels in white as background. The contour is extracted from the binary seal images by detecting the first and the last foreground pixels for every row and every column, shown in Fig.2 (*b*) and (*d*).

Given that circle is a special kind of ellipse, we indiscriminately fit the contour points of the two kinds of seal images for ellipse. An ellipse is a special case of a general conic which can be described by an implicit second order polynomial $F(x,y)=ax^2+bxy+cy^2+dx+ey+f=0$ with an ellipse-specific constraint $b^2-4ac<0$. The fitting of a general conic to a set of points $(x_i,y_i),i=1...N,$ can be approached by minimizing the sum of squared algebraic distances $\min_i \sum_{i=1}^{N} F(x_i,y_i)^2$. For this reason, a non-iterative ellipse-specific fitting method [9] can be used to get the coefficients a,b,c,d,e,f by which the semi-major axis A, the semi-minor axis B, the center (x_0,y_0), the angle θ_0 from the major axis to the X-axis in clockwise direction can be worked out too. So we can automatically classify an input seal image into either circular seals or elliptical ones by the ratio of A to B. If the ratio is larger than a threshold this seal imprint belongs to the elliptical shape. Otherwise it belongs to the circular one.

If a sample seal image is classified into circular shape category, the center (x_0,y_0) of the fitted ellipse plays as the center of the original circular seal image. And B is defined as the radius R of the original image. If the sample seal image is identified as the elliptical shape, it is necessary to rotate the original image by the angle θ_0 in clockwise direction so that the elliptic general equation exists, which is defined as $(x-x_0)^2/a^2+(y-y_0)^2/b^2=1$. After that, A and B do not change. And the center (x_0,y_0) needs to be updated and still represents the center of the ellipse seal image.

III. CHARACTER SEGMENTATION AND SKEW CORRECTION

Character segmentation and skew correction is described in detail in [4]. Generally speaking, the most import step is to use mathematical transformation to convert the circularly arranged or elliptically arranged characters into a rectangular region. Then segment the characters from this region.

A. Transforming the Sample Image into a Rectangular One

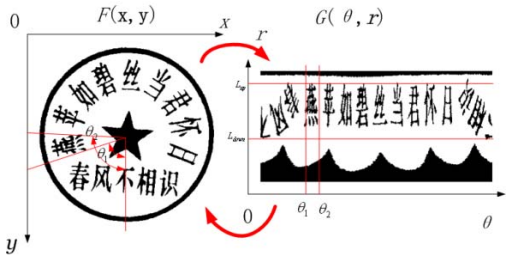


Figure 3. Transforming process for circular seal images

Fig.3 shows the process that the input circular seal $f(x,y)$ is transformed into a rectangular image $g(\theta,r)$. According to (1), the pixel of (θ,r) in a rectangular image $g(\theta,r)$ takes the same value as the pixel of (x,y) in the circular seal image $f(x,y)$. The created rectangular image is θ pixels' wide and r pixels' high ($r=0,...,R$， $\theta=0,...,359^\circ$).

$$x = x_0 + r\cos(\theta+90^\circ), y = y_0 + r\sin(\theta+90^\circ) \quad (1)$$

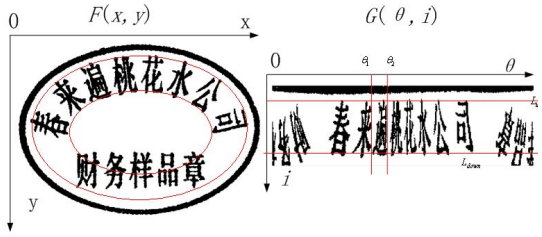


Figure 4. Transforming process for elliptical seal images

However, the elliptic general equation has two parameters, such as the semi-major axis A and semi-minor axis B, while the round general equation has only one such kind of parameter which is the radius R. So it is necessary to use a different way to handle elliptical seal imprints. As shown in Fig.4, according to (2), the pixel of (θ,i) in a rectangular image $g(\theta,i)$ takes the same value as the pixel of (x,y) in the elliptical seal image $f(x,y)$. The created rectangular image is θ pixels' wide and i pixels' high ($i=0,...,2/3B$， $\theta=0,...,359^\circ$).

$$\begin{aligned} x &= x_0 + (A-i)\cos(\theta+90^\circ) \\ y &= y_0 + (B-i)\sin(\theta+90^\circ) \end{aligned} \quad (2)$$

B. Character Segmentation for Circular Seal Images

It is necessary to remove the frame and the pentacle in the image $g(\theta,r)$. As shown in Fig.3 $g(\theta,r)$, there is always a gap between the main character region and the frame region and a gap between the main character region and the pentacle region. By traversing the rectangular image, we calculate the number of foreground pixels for every row, and then the rows with the minimum number of foreground pixels consist of two gaps. So the main character region is between the two gaps. And it is easy to get the upper boundary L_{up} and the lower boundary L_{down} of the main region. The next step is to change all the foreground pixels which are higher than L_{up} or lower than L_{down} to background.

The next step is to segment characters for the rectangular image $g(\theta,r)$ on the θ-axis and find every character's left and right boundaries. If the input seal image belongs to the category of specific-use circular seals as shown in Fig.1 (*b*), there are always the auxiliary parts on the left and right of the main region in the rectangular image $g(\theta,r)$. It is necessary to change these pixels to background. Meanwhile, we should also avoid splitting a Chinese character into two parts during the horizontal character segmentation process.

The whole horizontal segmentation process is as follows:

Step 1: Traverse the entire rectangle image and calculate the number of foreground pixels for every column. It is obvious that there is always a gap between two adjacent characters. Therefore, by calculating the column with the minimum number of foreground pixels, we can get every character's left boundary θ_1 and right boundary θ_2. And every character's width w on the θ-axis equals to $(\theta_2 - \theta_1)$.

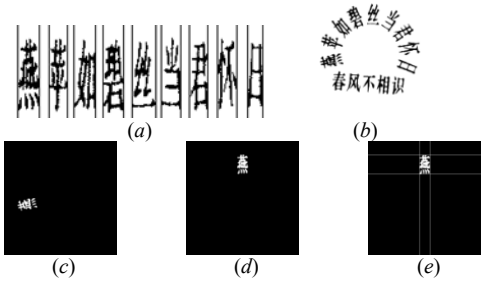
Step 2: Among all of these segments' w, get their median value w_{median} as a reference. Calculate the maximum width w_{max} and the minimum width w_{min} by multiplying w_{median} with 1.7 and 0.625 respectively.

Step 3: Compare every segment's w with w_{max} and w_{min}: ① If $w_{min} < w < w_{max}$, this segment meets the standard Chinese character requirement. ② If $w < w_{min}$, this segment is not a standard character and it is necessary to combine this segment with the next one to generate a character. And use this new w to compare with w_{max} and w_{min} again. ③ If $w > w_{max}$, this is a part of the auxiliary region. And change these pixels into background pixels as shown in Fig.5(*a*).

The next step is to remove the frame region and the pentacle region on the original round image $f(x,y)$. Calculate the distance d from every pixel (x,y) to the center (x_0,y_0). The distance d is defined as $d=((x-x_0)^2+(y-y_0)^2)^{1/2}$. If $d>L_{up}$, this pixel belongs to the frame region and it is changed into a background pixel. If $d<L_{down}$, this pixel belongs to the pentacle region and it is changed into a background pixel as shown in Fig.5 (*b*).

So far we have known every character's left boundary θ_1 and right boundary θ_2 on the rectangular image $g(\theta,r)$ which is also the angular span of the character relative to the Y-axis in clockwise direction in the original image $f(x,y)$. As shown in Fig.3, take the character "燕" for example. First, calculate the slope k of the line L which connects an arbitrary pixel (x,y) to the center (x_0,y_0), where k is defined as $k=(y-y_0)/(x-x_0)$. Then calculate the angle λ from the Y-axis to this line L in clockwise direction. If $\lambda<\theta_1$ or $\lambda>\theta_2$, change this pixel (x,y) into a background pixel, where there is only one character "燕" as shown in Fig.5 (*c*).

The next step is the skew correction stage. Since the skew angle which is the middle line of the character relative to the Y-axis in clockwise direction is $(\theta_1+\theta_2)/2+180^\circ$, it only needs to rotate the image $f(x,y)$ in counterclockwise by the angle $(\theta_1+\theta_2)/2+180^\circ$ so that the skew correction of this character is realized. Segment the character "燕" from the rotated image as shown in Fig.5 (*d*) and (*e*).



(*a*) (*b*) (*c*) (*d*) (*e*)

Figure 5. (*a*) Character segmentation on the θ-axis. (*b*) Remove the interference pixels. (*c*) Reserve only the character "燕". (*d*) Skew correction.(*e*) Character segmentation.

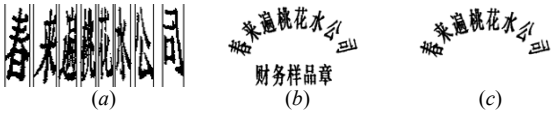
C. Character Segmentation for Elliptical Seal Images

The strategy of the character segmentation and skew correction for elliptical seal images is similar to that for circular seal images but with a little change. In the rectangular image $g(\theta,i)$ there is also a gap between the main characters region and the frame and a gap between the main region and the lower boundary of the image $g(\theta,i)$. So we use the same method to remove the frame region, achieve the character segmentation on the θ-axis of the rectangular image $g(\theta,i)$, remove the auxiliary region and find every character's left boundary θ_1 and right boundary θ_2, as shown in Fig.6 (*a*). Then detect the upper boundary L_{up} and lower boundary L_{down} of the main region as shown in Fig.4.

Based on the properties mentioned above, we can conclude the quadratic equations of the circumscribed ellipse and the inscribed ellipse of the main region in the original image $f(x,y)$, which is defined as $f_C(x,y)$ and $f_I(x,y)$ respectively in (3), as shown in Fig.7.

$$f_C(x,y)=\frac{(x-x_0)^2}{(A-L_{up})^2}+\frac{(y-y_0)^2}{(B-L_{up})^2}-1$$
$$f_I(x,y)=\frac{(x-x_0)^2}{(A-L_{down})^2}+\frac{(y-y_0)^2}{(B-L_{down})^2}-1 \quad (3)$$

The next step is to remove the frame on the original image $f(x,y)$. Calculate the distance d from every pixel (x,y) in the elliptical seal image $f(x,y)$ to the center (x_0,y_0), which is defined as $d=\sqrt{\frac{(x-x_0)^2}{(A-L_{up})^2}+\frac{(y-y_0)^2}{(B-L_{up})^2}}$. If $d>1$， this pixel belongs to the frame region and it is changed into a background pixel. Seeing that there is a gap between the main character region and the auxiliary region, we can remove the auxiliary region as shown in Fig.6 (*b*) and (*c*).



(*a*) (*b*) (*c*)

Figure 6. (*a*) Character segmentation on the θ-axis. (*b*) Remove the frame region. (*c*) Remove the auxiliary region.

Here, the method of the character segmentation and skew correction for elliptical seals is different from that for circular seals. Because we cannot decide whether a pixel (x,y) in the elliptical image $f(x,y)$ is in the angular span of the character relative to the Y-axis in the clockwise direction by calculating the slope of the line connected a pixel (x,y) with the center (x_0,y_0), we have to use a different strategy. Take the character "遍" as shown in Fig.8 for example. Since we have gained every character's left boundary θ_1 and right boundary θ_2 on the rectangle image $g(\theta,i)$, we apply them into (4) and gain 4 points on the circumscribed ellipse and the inscribed ellipse respectively as shown in Fig.7. Because two points determine a line, we can get the equation of the line connecting (x_{11},y_{11}) with (x_{12},y_{12}) which is $f_1(x,y)=(y_{12}-y_{11})(x-x_{11})-(x_{12}-x_{11})(y-y_{11})$, and that of the line connecting (x_{21},y_{21}) with (x_{22},y_{22}) which is $f_2(x,y)=(y_{22}-y_{21})(x-x_{21})-(x_{22}-x_{21})(y-y_{21})$. Then we apply each point (x,y) with $f_1(x,y)$ and $f_2(x,y)$ respectively. If $f_1(x,y)\geqslant 0$ and $f_2(x,y) \leqslant 0$, this pixel belongs to the area of this character and should be preserved. Otherwise it should be changed into a background pixel as shown in Fig.8 (*a*).

$$\begin{cases} (x_{11},y_{11})=\left(x_0+\left(A-L_{down}\right)\cos(\theta_1),y_0+\left(B-L_{down}\right)\cos(\theta_1)\right) \\ (x_{12},y_{12})=\left(x_0+\left(A-L_{up}\right)\cos(\theta_1),y_0+\left(B-L_{up}\right)\cos(\theta_1)\right) \\ (x_{21},y_{21})=\left(x_0+\left(A-L_{down}\right)\cos(\theta_2),y_0+\left(B-L_{down}\right)\cos(\theta_2)\right) \\ (x_{22},y_{22})=\left(x_0+\left(A-L_{up}\right)\cos(\theta_2),y_0+\left(B-L_{up}\right)\cos(\theta_2)\right) \end{cases} \quad (4)$$

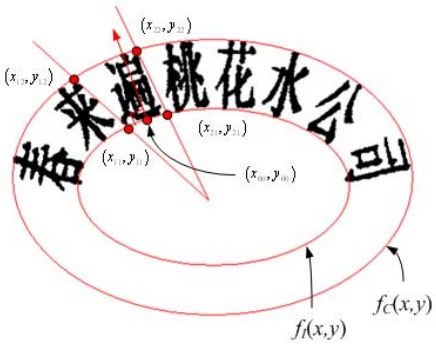


Figure 7. Schematic illustration of character skew correction

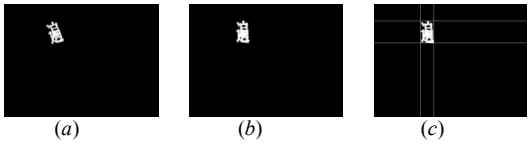
Then calculate the mean value θ_{00} of the angular span of the character "遍", which equals to $(\theta_1+\theta_2)/2$. And the coordinates (x_{00},y_{00}) of the point on the inscribed ellipse of the main region can be worked out as follows: $(x_{00},y_{00})=(x_0+(A-L_{down})cos(\theta_{00}), y_0+(B-L_{down})cos(\theta_{00}))$. Hence, it is obvious that the skew angle of every character equals to the normal direction of the point (x_{00},y_{00}) on the inscribed ellipse. The slope of the point (x_{00},y_{00}) on the inscribed ellipse can be worked out as follow formula:

$$\nabla f_{内}(x,y)=\left\{\frac{\partial f_{内}}{\partial x},\frac{\partial f_{内}}{\partial y}\right\}=\left\{\frac{2(x-x_0)}{(A-L_{down})^2},\frac{2(y-y_0)}{(B-L_{down})^2}\right\}. \quad (5)$$

Therefore, we can get the slope k as follows:

$$k=\frac{(y_{00}-y_0)(A-L_{down})^2}{(x_{00}-x_0)(B-L_{down})^2}. \quad (6)$$

Based on this parameter k, we can calculate the skew angle relative to the Y-axis in clockwise direction for every character. Just rotate the character in counterclockwise by this angle so that the skew correction of this character is realized. Segment the character "遍" from the rotated image as shown in Fig8. (*b*) and (*c*).



(*a*) (*b*) (*c*)

Figure 8. (*a*)Reserve only the character "遍". (*b*) Skew correction. (*c*) Character segmentation.

IV. OCR STRATEGY

Compared with standard machine printed form, the character images extracted from seals are always skew, noisy, and their strokes are always thicker. These differences have a great impact on the accuracy of the whole OCR system. It is very hard to use single OCR engine to get satisfactory recognition results.

In our OCR module, we combine three different engines with the aim of overcoming the limitations of individual classifiers. The classifier structures are as follows: OCR1 is a learning vector quantization (LVQ) classifier, OCR2 is a modified quadratic discriminant function (MQDF) classifier, OCR3 is a Euclidean distance (EUD) classifier. The features used by these three classifiers are gradient direction feature (GRD), chain code feature (CHF), and traversing-time feature (TRF) respectively.

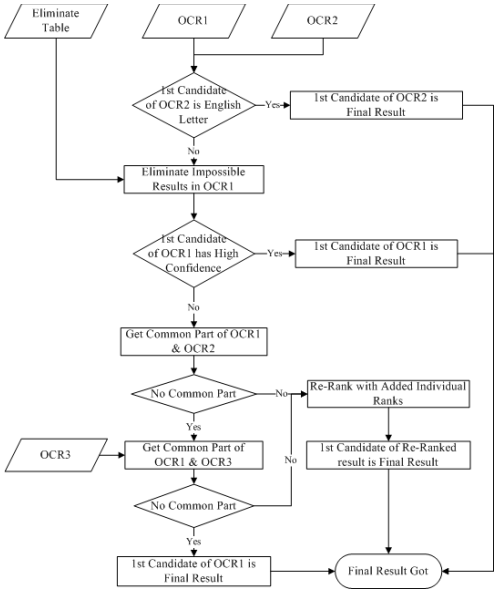


Figure 9. Workflow of our proposed OCR strategy

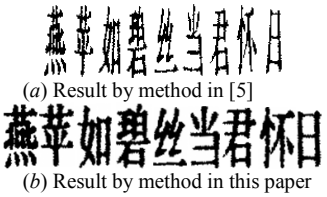
In our combination system, we depend mainly on the output of OCR1. For OCR1 can recognize more than 6000 characters, some of which will never appear in a formal seal. we eliminate some candidate results before combination. Then we examine OCR1 results in measurement level (class scores). If the confidence of first candidate is much higher than the second, we take the first candidate of OCR1 as final result. Otherwise we take the results of the other 2 engines into consideration and combine them with OCR1 in rank level (rank order). Common part is taken from the top few candidates of OCR1 and OCR2 first. And for each result in the common part, their ranks in both OCR engines are added as its final rank. Re-rank the common part and take the first one as the final result. If no common part shared between OCR1 and OCR2, OCR3 is used to combine with OCR1 in the same way. If common part is still not found (rarely happens), the first candidate of OCR1 is used as the final result. Additionally, we find that OCR2 performs much better on recognizing English letters. So if the first candidate of OCR2 is English letter, it is taken as the final result in our system. The flow chart of the OCR strategy is shown in Fig.9.

V. EXPERIMENTS

To verify the effectiveness of the algorithm described above, 663 seal images with different shapes are collected, including official circular seal images, specific-use circular

seals images and special-use elliptical seal images. The methods of Chinese character segmentation, skew correction and OCR strategy introduced above are applied to this seal image database. Consequently, 653 sample seal images are correctly segmented and the Chinese characters in these seal images are extracted and identified. The inaccurate results derived from poor quality of some sample seal images. Finally, the segmentation accuracy reaches 98.5%. Partial results are shown in Figs.10, 11, 12.

The comparison is shown in Fig.10 (*a*) and (*b*), Fig.11 (*a*) and (*b*). Fig.10 (*a*) and Fig.11 (*a*) show the results by using the method [5] to handle a specific-use circular seal image and an official circular seal image respectively. It is obvious that the characters segmented have severe distortion and lose a lot of detailed information. Meanwhile, Fig.10 (*b*) and Fig.11 (*b*) show the results by using our proposed method to handle the same seal images. Our results are better. In addition, the elliptical seal images which could not be addressed by the traditional methods such as [5] can be handle well by our proposed method as shown in Fig.12 (*a*) and (*b*).

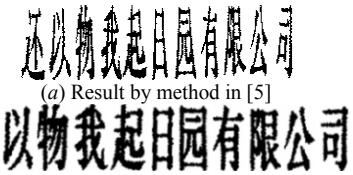


(*a*) Result by method in [5]

燕苹如碧丝当君怀日

(*b*) Result by method in this paper

Figure 10.

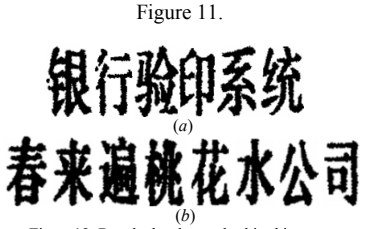


(*a*) Result by method in [5]

还以物我起日园有限公司

(*b*) Result by method in this paper

Figure 11.



(*a*)

(*b*)

Figure12. Results by the method in this paper

Table Ⅰ shows the recognition rates of individual classifiers and their combination. It can be easily seen that the fusion of the three classifiers yields higher performance and reaches 97.0% recognition rate. It also shows that the three classifiers differing in feature representation, architecture and training data exhibit complementary classification behavior in recognizing extracted Chinese seal characters.

TABLE I. RECOGNITION RATES (%) OF INDIVIDUAL CLASSIFIERS AND THE COMBINED SUPER CLASSIFIER

OCR Engine	Feature	Classifier	Accuracy
OCR1	GRD	LVQ	88.9%
OCR2	CHF	MQDF	68.0%
OCR3	TRF	EUD	72.3%
Combined			97.0%

VI. CONCLUTION

This paper has proposed a new method for automatic Chinese seal identification using a novel character segmentation method on Chinese seal images and a combined OCR strategy. By incorporating the characteristics of elliptical and circular seals, we can segment Chinese characters from both circular seal images and elliptical seal images. To overcome the limitation of individual OCR classifier, a combined scheme using three different OCR classifiers is developed and it improves the recognition accuracy significantly. Experiments on the collected seal images have demonstrated that our proposed method works well for seal retrieval.

REFERENCES

[1] K. Ueda, K. Matsuo, "Automatic seal imprint verification system for bank check processing", Third International Conference on Information Technology and Applications (ICITA 2005), Jul. 2005, Vol.1, pp.768-771, doi:10.1109/ICITA.2005.81.

[2] P.P Roy, U. Pal, J. Llados, "Seal detection and recognition: an approach for document indexing", 10th International Conference on Document Analysis and Recognition (ICDAR 2009), Jul. 2009, pp. 101 -105, doi:10.1109/ICDAR.2009.128.

[3] K. Ueda, "Extraction of signature and seal imprint from bank checks by using color information", Third International Conference on Document Analyisis and Recognition (ICDAR 1995), Aug. 1995, Vol.2, pp.665-668, doi:10.1109/ICDAR.1995.601983.

[4] Chao Ren, Youbin Chen, "A new method for character segmentation and skew correction on Chinese seal images", 2nd World Congress on Computer Science and Information Engineering (CSIE 2011), Jun. 2011, in press.

[5] Hong Liu, Ye Lu, Qi Wu and Hongbin Zha, "Automatic seal image retrieval method by using shape features of Chinese characters", ISIC. IEEE International Conference on Systems, Man and Cyberntics (ICSMC 2007), Oct. 2007, pp.2871-2876, doi: 10.1109/ ICSMC. 2007. 4414010.

[6] Xiaoyang Wang, Youbin Chen, "Seal image registration based on shape and layout characteristics", 2nd International Congress on Image and Signal Processing (CISP 2009), Oct. 2009, Vol.7, pp.3440-3444, doi: 10.1109/CISP.2009.5302120.

[7] Yuanfei Cheng,"Seal recognition using the shape selection algorithm", IEEE International Conference on Electro/information Technology, (EIT 2006), May. 2006, pp.544–547, doi: 10.1109/EIT.2006.252207.

[8] The State Council of the P.R.China, "Provisions of the State Council on the management of seals used by state administrative organs, businesses and public institutions and social organizations", the Sstate Council issued, 1999, No.25.

[9] R. Halif and J. Flusser, "Numerically Stable Direct Least Squares Fitting of Ellipses", Department of Software Engineering, Charles University, Czech Republic, 2000.