Automatic Seal Image Retrieval Method by Using Shape Features of Chinese Characters

Hong Liu, Ye Lu, Qi Wu and Hongbin Zha
State Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School
{liuhong, luye, wuqi, zha}@cis.pku.edu.cn

***Abstract*—In many eastern countries, a large number of seal images need to be identified every day. The system compares an input seal image with its reference seal and validates the authenticity of it. However, the reference seal is usually found manually. As each reference seal has quite a long corresponding ID, inputting ID to get the needed seal one by one costs quite a lot of time and this manual stage has become the bottleneck of automatic seal identification. To make the seal verification system more automatic, a new retrieval method based on Chinese characters' shape features will be introduced in this paper. Firstly, the main characters region is obtained by transforming the round seal image to a rectangle one and choosing the main part for each seal. Secondly, every single character is segmented using correlation method, and the number of characters in a seal can be got. Thirdly, four horizontal and four vertical features are extracted for each character in a seal and an eigenvector called position code is defined. Finally, the retrieval strategy is mainly based on transform from position code to a weight to decide which two seal images have the most similarity. Experiments on a database of 1000 testing seal images provide the retrieval ability for the proposed approach. The correct rate goes to 95.3%.**

I. INTRODUCTION

In China and some other Asian countries, seal image has been commonly used for personal confirmation. With the rapid development of modern science and technology, the traditional manual management becomes difficult to meet the requirement of seal identification. To identify seal images' authenticity by computer automatically becomes a subject with great theoretical significance and application value.

So far, there has been some research on seal image processing, especially for the verification of it. One method does not deal with the problem as a general pattern matching problem. Instead, two kinds of costs called the negative cost and the positive one which stand for the correlation between an input image and the reference seal are defined [1]. Some research presents an application of the fuzzy integral for the selective extraction of color clusters in computer vision systems. The approach is applied on document analysis for the isolation of seals [2]. Another method is proposed for rotation invariant verification, considering that a seal image is probably skew in a document image [3] [4]. Reference [3] verifies seal image by using the discrete K-L expansion of the DCT after three steps. Reference [4] represents the rotation invariant features by the coefficients of 2D Fourier series expansion of the log-polar image. A comprehensive method has already been specifically designed for application in Japanese bank check processing [5] [6]. The method combines two different algorithms for seal image verification. The first verification algorithm is based on a method using local and global features of seal image. The second algorithm uses a special correlation based on a global approach.

For all of the methods talked above, seal image processing has been realized on traditional seal identification systems. In such a system, an index is always established, which means every reference seal has a particular ID. When an image is input, the relevant ID is input manually so that the correct reference seal can be called. Then the verification stage is carried out, using methods quoted in the last paragraph. Meanwhile, the automatic identification system introduced in this paper establishes an index according to the content of seal image instead of ID, which leaves out the manual stage of inputting the correct ID. Whenever a seal image is input, the system can find the corresponding reference seal in the database by the features shown by it. It is obvious that the new automatic system is more intelligent than the traditional one.

The OCR method has also been widely used for retrieval, but it does not fit the purpose of seal image processing. OCR identifies every character one by one based on stoke information. But seal images are often vague so that OCR is hard to identify them well. However, the method proposed in this paper is based on the structure information of a whole seal and the profile information of each character so it is more robust on seal image retrieval.

In this paper, a method on how to automatically find the correct reference seal for the input image before the verification stage is presented. The method uses the information of positions instead of lengths of features so that it will be robust dealing with vague seal images. Although there are lots of different fonts for Chinese characters, which add much trouble to seal image processing, the method is also effective among them. It is a rotation invariant method, too.

In the following sections, Section Ⅱ and Section Ⅲ introduce the process on how to get the properties and features of a seal. The experiment results of this new method

1-4244-0991-8/07/$25.00/©2007 IEEE

are given in Section Ⅳ and conclusions are presented in Section Ⅴ.

II. EXTRACTION OF MAIN CHARACTERS REGION

Fig.1 shows a check with seal images on it. There are over ten million such checks which need to be identified in China every day. Our laboratory is building a seal image processing system, which includes the stages of location, extraction, retrieval and identification.

The stage of location and extraction is also a difficult problem to deal with. Since the method in this paper just focuses on the stage of retrieval. Only a brief introduction is given here.

The first step is to transform a check to a binary image. A local Laplacian method is used here. The second step is to extract the seal from the complex image. After wiping off noisy pixels, long horizontal beelines are discarded using the information of foreground pixels' run-length. Thus seals can be independent connective regions. At last, the algorithm signs these connective regions which are in proper size.

Fig. 1. A check with seal imprints

The seals in the database used in this paper are of a typical type in China, which all have round figure and their common features. Two of the seals are shown in Fig. 2. The seals have been changed to binary images from RGB images, which is the pretreatment stage of the method. The main characters surround the pentacle in the center of round seal so that it is convenient to extract them. The seals can be classified into two categories. In Fig.2, (a) represents the seals in which not all of the characters belong to the main region. In fact, some of the characters belong to the auxiliary region and they will not be used in the method proposed in this paper (These characters are in a line instead of along an arc.). They are defined as Category 1. And in Fig.2, (b) represents the seals in which all of the characters belong to the main region. They are defined as Category 2. In all binary images in this paper, pixels in black are defined as foreground pixels and pixels in white are defined as background pixels.

After finding the center of a seal by statistical method, the seal with round figure is transformed into a rectangle image retaining the part in which all the main characters are included.

The transformation is defined as follows:

$$x = x_0 + [r_1 + (r_2 - r_1)/n \times k] \times \sin\theta$$
$$y = y_0 + [r_1 + (r_2 - r_1)/n \times k] \times \cos\theta \quad (1)$$

(a) Category 1 **(b) Category 2**

Fig. 2. Binary seal images

The rectangle image created is m pixels' wide and n pixels' high (k=0,1,2,…,n-1; θ=0,0.5,1,1.5,2,2.5,…,359.5). (x_0,y_0) are the coordinates of the center of round seal and r is the radius of it. Here r_1 equals r/2 and r_2 equals r.

As shown in Fig.3, through (1), the pixel of (k,θ) in a rectangle image takes the same value as the pixel of (x,y) in the round seal image. Then the whole rectangle image is created as shown in Fig.4.

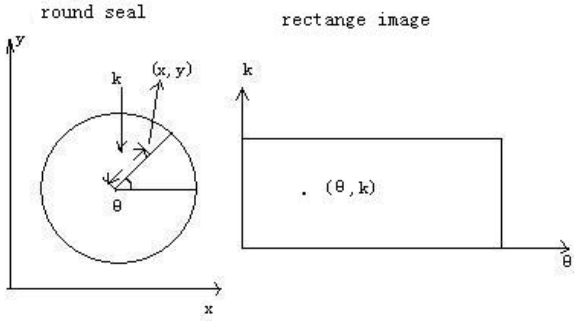
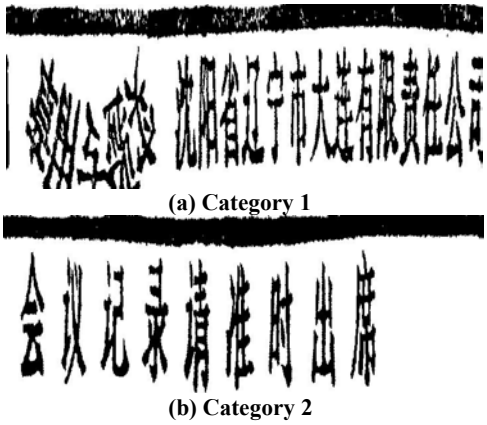


Fig. 3. Transformation from a round seal to a rectangle image



(a) Category 1

(b) Category 2

Fig. 4. The rectangle image

It is obvious that the rectangle image in Fig.4 can not be directly used for the segmentation stage. It is necessary to add an extracting effective part stage, which includes two aspects, the vertical and horizontal ones.

Vertically, the major objective is to remove the interference pixels in the higher half, which is shown in Fig.4. Because there is always a gap between the main characters region and the interference pixels, it is easy to realize it by traversing the higher half of a rectangular image

and calculating the number of foreground pixels for every row. The row with the minimum number of foreground pixels is defined as the sign row.

The next step is to change all the foreground pixels which are higher than the sign row to background pixels and the vertical stage is done.

As shown in Fig.4, horizontally, the main characters region is between the two widest gaps for Category 1, and the region is from the widest gap's tail to its head for Category 2.

The whole horizontal process is as follows:

Step 1: The system traverses the entire image column by column. If any of the columns has no foreground pixel, mark it as the blank column (after the vertical stage).

Step 2: It combines together all the blank columns which are next to each other one by one, so that a group of gaps are obtained. It also finds the first and second widest gaps, which is respectively defined as gap *a* and gap *b*.

Step 3: If the margin between the widths of *a* and *b* is smaller than a threshold *t*, the seal belongs to Category 1. Then it goes to Step 4. Otherwise, the seal belongs to Category 2, and it goes to Step 5.

Step 4: There are two parts between *a* and *b*, and the longer one is the main characters region for Category 1. Then the region is got. It goes to Step 6.

Step 5: The system needs to find the main characters region only using *a*, which is from gap *a*'s tail to its head. Then the region is got.

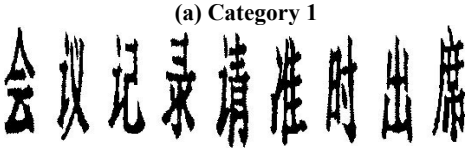
Step 6: Save the main characters region for the next character segmentation stage.

Step 7: End.

Although the round figure seal has already been changed to a rectangle image, it should always be considered that the head and the tail of this rectangle are born connected. It means whenever it is necessary, the tail should be disposed followed by the head.

After all the stages introduced in section Ⅱ, the main characters region can be got as shown in Fig.5.

(a) Category 1



(b) Category 2

Fig. 5. The main characters region

III. CHARACTER SEGMENTATION AND FEATURE EXTRACTION

For Chinese seals, to extract the features from a single character is more convenient than from the whole main characters region, so there is still a character segmentation stage before the feature extraction stage.

The segmentation of characters can be obtained based on the method of correlation. The correlation formula is shown as follows.

$$f(j)=\sum_{i=1}^{l/2} a(i)\times a(i+j) \tag{2}$$

In formula (2), *i* (*i*=1,2,…,*l*) stands for the serial number of every column in a main characters region image, where *l* is the total number of columns. And *a(i)* stands for the number of foreground pixels for every column. Here *f(j)* is the result of correlation transform. Function *f(j)* is shown in Fig.6, where *j*=1,2,3,…,*l*/2. Based on the periodicity of the number of foreground pixels in each column, which is produced by characters one by one, the cycle can be got by calculating the horizontal distance between two wave crests which are next to each other in Fig.6. As the cycle turns to be just the average width of a character, characters segmented image can be got by segmenting the image every cycle pixels, which is shown in Fig.7 as the vertical lines.

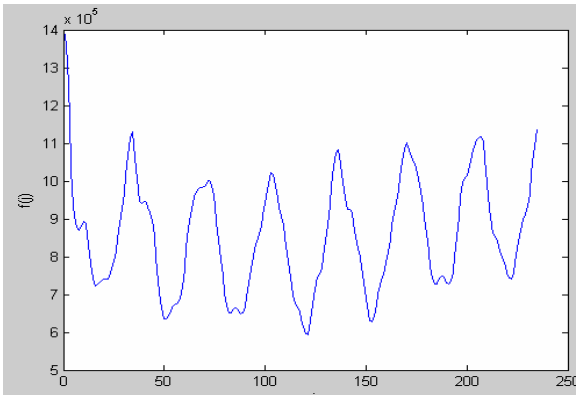


Fig. 6. Correlation function f(j)

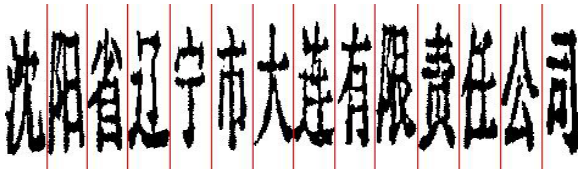


Fig. 7. The characters segmented image

However, there are still some exceptions. The entire process is shown in Fig.8, considering these exceptions.

After the character segmentation stage, the feature extraction stage should be carried out. There are 4 features respectively on both horizontal and vertical directions for a character (totally 8 features), which are a row or a column with the first largest, the second largest, the second smallest and the first smallest number of foreground pixels. In both directions, they are respectively defined as shoulder, crotch, shank and waist of a Chinese character compared with a human being.

These eight features are all indispensable. Both horizontal and vertical features are extracted so that more

comprehensive information can be got. In each direction, the four features all have their own necessity. Because the two widest ones are the most stable information and the two narrowest ones are most representative. Also, two but one widest features are chosen because the first and the second widest ones are easy to be confusable. The two narrowest ones may also be in the same occasion.

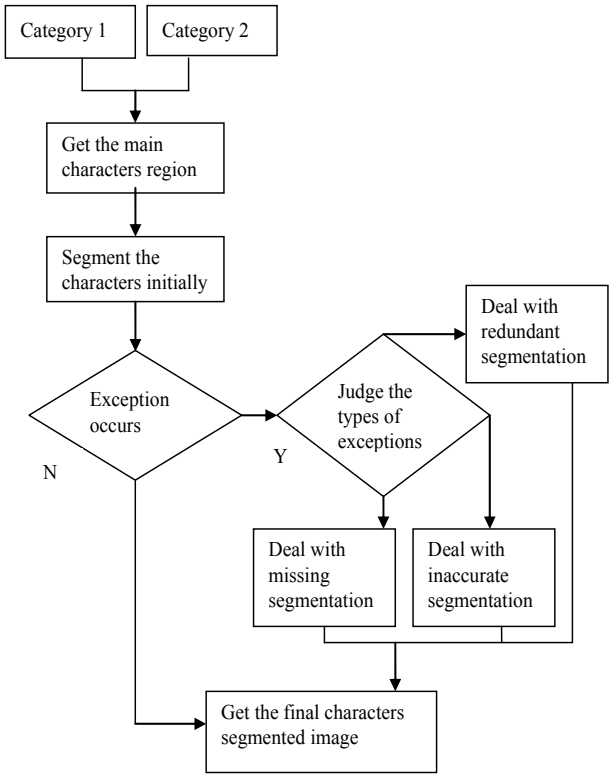


Fig.8. Entire process of character segmentation

The features of a character are shown in Fig.9. The Chinese Character Liao is extracted from the seal in Fig.2 (a) after all the stages introduced above and the shape features are presented. The narrow lines stand for horizontal features and the wide lines stand for vertical ones. In both directions, the two longer lines are the shoulder and the crotch, and the two shorter ones are the shank and the waist. The shoulder and the crotch should not be too close in case of losing some essential features of the character. It is the same between the shank and the waist, and they are also not allowed to be too close to the borderlines as Chinese characters often start or end with a narrow stoke.

The positions of 8 features for each character are defined as a position code *pc*.

$$pc=(p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8) \quad (3)$$

p_1, p_2, p_3, p_4, p_5, p_6, p_7 and p_8 are the positions where the relevant features appear, which are expressed by the serial number of them.

The reason why positions but the lengths of features are recorded is that seal images are often vague and positions are much robuster than lengths of features. Also, it can save a large amount of time by using information from positions instead of lengths.

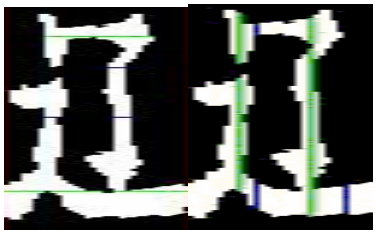


Fig. 9. Features of a character

To transform position code to the similarity of two characters, a weight is defined in formula (4) and formula (5) to measure the degree of similarity of a pair of seals which have the same number of main characters. The pair of characters which have larger w_c means they are more similar to each other.

$$w_c = \sum_{i=1}^{8} w_{fi} \quad (4)$$

$$w_{fi} = \begin{cases} a_1 & (|p_{1i} - p_{2i}| < b_1) \\ a_2 & (b_1 < |p_{1i} - p_{2i}| < b_2) \\ a_3 & (b_2 < |p_{1i} - p_{2i}| < b_3) \\ 0 & (|p_{1i} - p_{2i}| > b_3) \end{cases} \quad (5)$$

Here w_c stands for the weight of two characters which are ready to be compared, and w_{fi} is the weight of one of the eight features. The weight of any feature can be classified into three categories, where w_{fi} equals a_1, a_2, or a_3 respectively ($a_1>a_2>a_3$). And p_{ji} (j=1,2; i=1,2…8) are extracted from position codes, where i stands for different seals and j stands for different features. Here b_1, b_2 and b_3 are experimental boundary values ($b_1<b_2<b_3$).

The weight of each pair of seals which have the same number of characters should be the summation of all the w_c values from each character.

As shown in formula (5), when a pair of characters' corresponding features appear at closer positions, their weight should be added a larger value. When they appear at farther positions, their weight should be added a smaller value. But when they appear too far from each other, their weight should just be immovable. There should not be any subtraction of weight, as it is probably caused by a noise.

There are only about 2500 Chinese characters which appear in high frequency. Up to 99% characters in common use are from these 2500 characters. Classified by different sequences of position codes, all of Chinese Characters can be put in 576 groups (24 multiplied by 24). So there would be not more than 5 characters in each group on average, on the assumption that each group had more or less the same number of characters. The combination of characters in a seal is restricted by both syntax and meaning, so there are very few chances that a pair of different characters are very close to each other based on shape features.

IV. EXPERIMENTS

Using the algorithm mentioned in the last two sections, the width of main characters region, the number of characters and the position code for each seal can be obtained. A database with 1000 seal images made from 50 different reference seals is established. In the database, each seal has its own name, serial number, width of main characters region and the number of characters. Moreover, position codes are also recorded. Table.1 shows the information for a seal in the databse.

Table. 1. Information for a seal in the database

name	11-1.jpg							
serial number	1							
width	444							
characters number	12							
position code	237	89	192	142	225	89	173	143
	184	81	244	144	160	145	115	91
	238	99	85	137	112	135	145	84
	178	81	240	139	248	82	134	113
	145	78	241	111	245	101	172	223
	240	115	94	146	97	123	245	154
	13	30	5	12	45	59	54	48
	83	103	99	77	126	140	137	133
	172	181	159	152	202	190	213	198
	233	227	240	247	282	265	276	280
	319	304	331	329	352	342	358	369
	394	377	401	383	423	434	417	428

After building the database of all the reference seals, the retrieval process can be started. When an image is input, the corresponding reference seal should be found out correctly. Every image is made from one of the reference seals but not the same image that is used to build the database. The process of retrieval of the correct reference seal is shown in Fig.10.

Both the images which were used to build the database and the ones which were later input to test the validity of algorithm have random directions, which will not affect the validity of the system because the algorithm is a rotation invariant method. All of the rotation problems have been settled in extraction of main characters region stage.

In the experiment, m equals 720 pixels, n equals 300 pixels and threshold t is defined as 15 pixels. Parameter a_1, a_2, and a_3 are respectively defined as 4, 3 and 1. The experiment shows the method can find the corresponding reference seals for input images well. The correct rate goes to 95.3% for 1000 input seals. Fig.11 shows two groups of results. Each group includes a seal image and its corresponding reference seal found by the algorithm proposed in this paper.

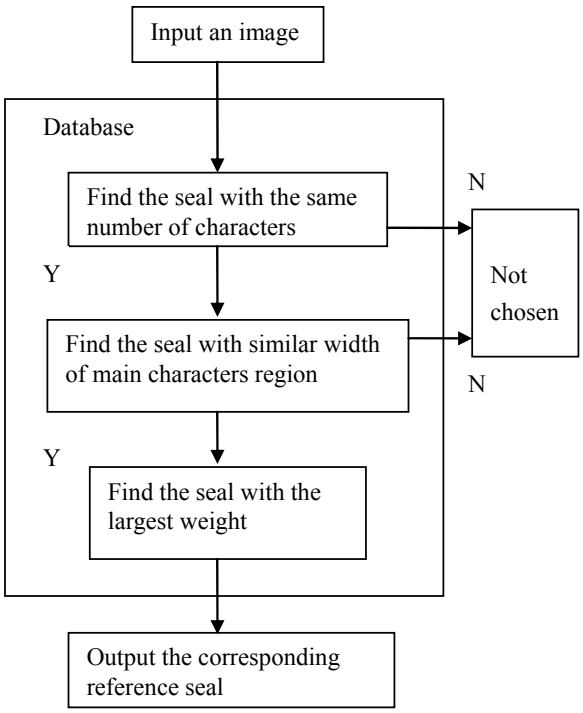


Fig. 10. Flow of retrieval for the correct reference seal

(a) Group 1

(b) Group 2

Fig. 11. Experiment results

In Fig.11, the two groups are from two different categories. In each group, the left one is the input image, and the right one is the reference seal selected from 50 different seal images in the database. The two groups both turn out to match correctly. It can be seen that the input image and the seal image that is used to build the database need not be the same one. Result will not be affected by different directions.

In Fig.11, the seal on the left in group 1 has 14 main characters. Unfortunately, there are 16 reference seals out of the 50 ones in database which all have the same number of main characters. Fig.12 shows the relation of matching weights between each of these 16 reference seals and the seal

on the left in group 1. In the graph, the abscissa stands for the serial number of each candidate reference seal, and the ordinate stands for the matching weight for each of them. It is clear that the No. 7 reference seal overwhelms the others. With the information from Fig.12 and also considering the width of main characters region as a consult, it is easy to select No. 7 as the corresponding seal. And in the database, No. 7 is just the seal image on the right in Fig. 11.

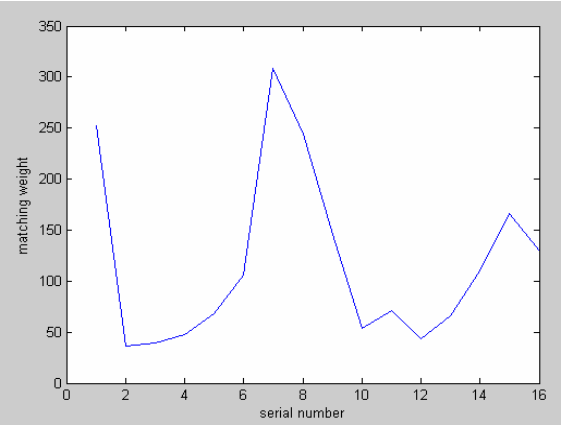


Fig. 12. Relation of matching weights

It is more convenient to select the corresponding reference seal for the seal image on the left in group 2. The seal has 9 main characters and there is only one reference seal in the database which has the same number of main characters. As a result, the seal on the right in group 2 is quickly selected without stage of calculating weights.

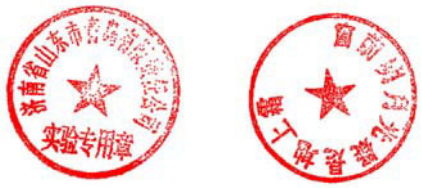


Fig. 13. Vague seal images

Many of the seal images are vague because of the problem of printing ink or the difference of strengths when being stamped. They can also be found out correctly by our method, as shown in Fig.13. But OCR is hard to realize it. The reason is that our method is based on shape features instead of each stroke of characters. Vague seal images which are so severe are called in-completed ones. From the procedure of the whole algorithm, it can be concluded that if the seal images is not so in-completed that even not enough information can be given to find the center of the seal correctly, the following steps will not be affected by the incompleteness badly.

V. CONCLUSIONS

This paper proposes an automatic seal image retrieval method based on Chinese characters' shape features. After the stage of extracting main characters region and the stage of extracting all characters' features, position code *pc* is defined. Then a weight is obtained from position code to test the similarity between two seals which have the same number of characters. A database with 50 samples is established, 1000 input images are tested and the correct rate goes to 95.3%. The matching result shows that the method works well and will not be affected by different directions of seal images. It can deal with vague seal images, too. The method can be combined with the methods of classical verification to realize a complete automatic system for seal identification.

Although the algorithm is only based on round Chinese characters now, it can be conferred that other characters also have their own shape features and for rectangle seals, characters are easier to extract. Then to find algorithms to search among seals in other languages depending on their particular shape features and to deal with seals of ellipse, diamond or other shapes are significant and challenging tasks. Our future work will focus on these fields.

Acknowledgements. This work is supported by Shenzhen Bureau of Science Technology and Information, National Natural Science Foundation of China (NSFC, No.60675025) and National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247).

REFERENCES

[1] T Horiuchi, "Automatic seal verification by evaluating positive cost", Proc 6th International Conference on Document Analysis and Recognition, 2001, pp.572-576.

[2] Aitreli Soria-Frisch, "The fuzzy integral for color seal segmentation on document images", ICIP, 2003, pp.157-160.

[3] T.matsuura, K.mori, "Rotation Invariant Seal Imprint Verification Method", ICECS, Vo1.3, 2002, pp.955-958.

[4] T.Matsuura, K.Yamazakiv, "Seal imprint verification with rotation invariance", Circuits and Systems, 2004, pp.597-600.

[5] K. Ueda, T. Mutoh, K. Matsuo, "Automatic verification system for seal imprints on Japanese bank checks", Proceedings of the 14th ICPR, 1998, pp.629-632.

[6] Katsuhiko Ueda, Ken'ichi Matsuo, "Automatic Seal Imprint Verification System for Bank check Processing", ICITA (1), 2005, pp.768-771.

[7] Y.-S. Chen, "Automatic identification for a Chinese seal image", Pattern Recognition, vo1.29, no.11, 1996, pp.1807-1820.

[8] Aureli Soria-Frisch., "Soft data fusion in image processing", in Soft-Computing and Industry: Recent Advances, R. Roy et al., Ed. 2002, Springer-Verlag, pp.4234.

[9] Javier Traver, Filiberto Pla, "Dealing With 2D translation estimation in log-polar imagery", Image and Vision Computing 21, 2003, pp.145-160

[10] Satoru OD0 and kiyoshi HOSHINO, "Estimation of target's shape and position from a monocular image sequence based on visual characteristics", technical report of IEICE HIP, 2001-84(2002-1).

[11] K. Ueda, H. Maegawa and K. Matsuo, "Automatic extraction of filled-in items from bank-check images", Int. Workshop on Document Analysis Systems, 2004, pp.225-228.

[12] S. Impedovo, P.S. Wang, and H. Banke Eds., "Special issues on Automatic bank check processing Part I and II", Int. J. Pattern Recognition and Artificial Intelligence, Vol.11, 1997.