Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Restoring Chinese Documents Images Based on Text Boundary Lines

Hong Liu
Key Laboratory of Machine Perception and Intelligence,
Key Laboratory of Integrated Microsystem,
Shenzhen Graduate School, Peking University, China
e-mail: hongliu@pku.edu.cn

Runwei Ding
Key Laboratory of Integrated Microsystem,
Key Laboratory of Machine Perception and Intelligence,
Shenzhen Graduate School, Peking University, China
e-mail: dingrw06317@szcie.pku.edu.cn

***Abstract*—Distortion always appears in document images while scanning thick bound volumes. There are two kinds of distortion for the scanned grayscale images, shadow appears at the volumes' spine area, and warping of the words occurs in the shadow. In this paper, a novel text boundary lines based method for efficient restoration of warped scanning Chinese document images is presented. We first detect on which side of an image the shadow lays by row grayscale analysis method. Then the shadow is removed by a modified Niblack's algorithm. In order to detect the warped feature, a text boundary lines' detection method is proposed. Finally, an adjustment method based on the text boundary lines is carried to restore the warped words. Experiments on 400 various scanning Chinese document images are implemented. The improvement on average character recall is 11.92% to 14.89%. Experiments show that the proposed restoration method is efficient for Chinese documents with both text and non-text regions.**

***Keywords*—distortion, text bounary lines, warped document images, restoration**

I. INTRODUCTION

With the development of office automation, more and more traditional volumes need to be transformed into electronic documents. Therefore, scanners are widely applied. However, scanning thick bound volumes always leads to distortions, shadow and warping near the spine of the volume which not only diminished document's readability, but also reduces the accuracy of OCR application. Therefore, the recovery of document images scanned from thick bound volumes is necessary for both human reading and text retrieval. In current literatures, techniques have been proposed for restoring warped document images. Those generally can be classified in two categories, one is 3D shape reconstruction based techniques, and the other is 2D image processing based techniques.

Methods of the first category have been reported by many authors. Kanungo et al. [1] introduced a global degradation model for perspective distortion. However, it assumed that the warped surface is circular cylindrical and the lighting direction is vertical. Wada et al. [2] developed a complicated model to reconstruct a 3D book surface, which assumes the surface is cylindrical and requires the spine of book to be patrolled to the scanning light. Complicated computation of the interreflection made the method having high computational cost. Pilu[3] proposed a method that models the surface of book using an applicable surface and restored the document image by unrolling the applicable surface to a plane. Brown and Seales [4] presented an approach to restore the warped document images using a mass spring model and a particle-based system. However, the shadow in the document images was not removed. Yamashita et al. [5] proposed an approach using a NURBS curve for reconstructing 3D surface. However, their method needed a two-camera vision system. It seemed that the 3D shape model based methods always need special setup and have lots of parameters which result in increased computation complexity, what is more, some of the parameters are difficult to be known in reality.

There are also several authors who proposed the techniques based on 2D image processing. Tang and Suen [6] proposed a method that approximates 3D distortion by using 2D geometrical transformations. Zhang et al. [7] adopted a regression model of curved text lines to restore the warped text lines. This method has a limitation that the word, especially long word, may be still warped after restoration. Catos et al. [8] presented a novel technique based on text lines and word detection. They recovered the document image word by word according to the word boundary lines. The approach of Ulges et al. [9] is based on line-by-line de-warping of the observed volume surface. They enclosed each letter with a quadrilateral cell, which is mapped to a rectangle of correct size and location in the result image. The existing 2D image processing based methods are usually used in document that only have text regions. If there are non-text regions in the documents, these 2D methods always fail. Although Zhong et al. [10] used the page boundary to correct the warped document, yet the whole page boundaries were difficult to be acquired by the scanner that limited its application.

This paper aims to propose a method which can correct warped Chinese document images, especially the documents which have both text regions and non-text regions without special setups and page boundaries. There are several problems left for complex document images. First, distinguishing text from non-text regions is rather difficult for warped documents. Second, computational complexity is very high to detect the warped features line by line even word by word. Third, the non-text regions warped feature is difficulty to be acquired without special setups. To solve the above problems a new warped feature should be found. From the warped documents

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

we found that the top and bottom lines of the document reflect the warped degree commendably. Therefore, a new concept of text boundary lines is proposed and a fast boundary lines' detection method of Chinese documents images is presented in this paper. Considering sometimes text boundary lines can not be acquired correctly while the non-text region is in the top or bottom of the documents, the similarity of the same volume side's adjacent pages is used to detect the boundary lines approximately. Then the documents can be restored based on the text boundary lines.

The remaining of this paper is organized as follows. The pretreatment will be introduced in section II. In section III, the upper and lower text boundary lines acquirement will be described. The restoration will be introduced in section IV. Experimental results and discussions will be given in section V. Finally, the paper will be concluded in section VI.

II. PREPROCESSING

In our method, the scanned document images should be preprocessed in order to get some information before its restoration, which is important and necessary for restoration work. The reason is that only when the warped side is known and the image is binaried, the text boundary lines could be acquired correctly.

A. Warped Side Detection

While scanning thick bound volumes, the obtained grayscale images will suffer from variant brightness and blurring. As shown in Fig. 1, shadow exists along the spine of the volumes.

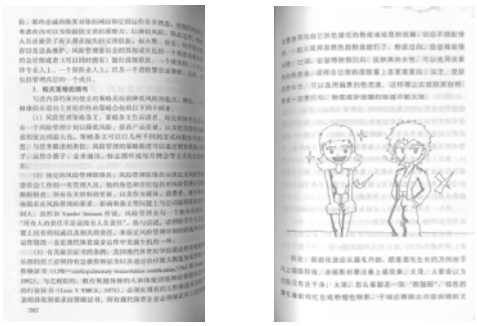

Figure 1. Example of image scanned from a thick

The shadow lies either on the left side or the right side of the images and only the warped side of the images suffers wrapping. Therefore, the warped side should be firstly detected using a row grayscale analysis function *V(x)*, which can be expressed as,

$$V(x_i) = \sum_{j=0}^{j=H} s(i, j) \tag{1}$$

$$s(i,j) = \begin{cases} 1 & s(i,j) < T \\ 0 & otherwise \end{cases} \tag{2}$$

here H is the height of the document image, $s(i, j)$ is the intensity value of pixel (x, y), and T is a predefined threshold. Comparing the local mean of the left side with the right side, the shadow lies in the bigger side. Fig. 2 shows the row grayscale analysis diagram of Fig. 1, the warped side can be easily detected from the diagram.

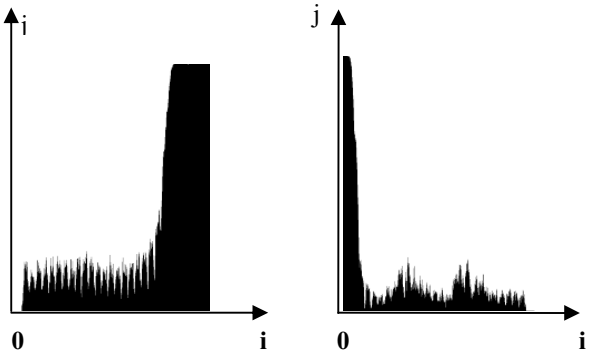

Figure 2. Row grayscale analysis of Fig. 1.

B. Modified Version of Niblack's Algorithm

The main idea of Niblack's algorithm [11] is to vary the threshold over the image. The threshold for each pixel $p(x, y)$ is computed from,

$$T(x, y) = m(x, y) + k * s(x, y) \tag{3}$$

where $m(x, y)$ and $s(x, y)$ are the local mean and local standard deviation calculated in a window which is a small neighborhood of the pixel $p(x, y)$. Addition, the window is normally uses 15*15. k is a user defined parameter which is negative in value. Niblack's algorithm is sensitive to the value of the k, therefore, it is difficult to find a k value that could produces best results for all experimental environments.

In order to remove the shadow of the warped document image, Z. Zhang and C.L. Tan [7] modified the Niblack's algorithm. In order to reduce the sensitivity of parameter k, $s(x, y)$ is divided by the dynamic range of standard deviation R, and $m(x, y)$ is utilized to multiply the standard deviation terms. Formula (4) presents the algorithm,

$$T(x, y) = m(x, y) \cdot [1 + k(1 - \frac{s(x, y)}{R})] \tag{4}$$

where R=100 and k =0.1.

In this paper, the modified Niblack's algorithm can not adopted directly. The reason is that the k value is not robust. It makes much noise in the shadow area, even results difficulties for our upper and Lower text boundary lines' acquirement. Therefore, the value of k should be redefined in our experiments.

C. Shadow Removal

In order to remove shadow, a modified Niblack's algorithm is used. In (4), k is a user defined parameter. Zhang et al. [7] used k=0.1 for grayscale images in order to get more information of the words in the document. Therefore, a filter is used to remove the noise. The static value limits their use in different experiments. Therefore, we modify k=0.05 and obtain better results. Fig. 3 shows the binarization result of Fig. 1.

In order to enhance the binaried result, a projection profile based method is applied to analyze the document image and remove the boundary of the page. In order to get the text boundary lines, the top or the bottom lines should be cut out if it is short line. As shown in Fig. 3, a short line lies at the bottom of the left page, the short line in this document is used to direct paginal number. After the short lines cutting out, we get the main body of the document. Sometimes, the volume is not perpendicular to the lens of scanner. Our group's skew detection method using robust borderlines [12] is used to adjust the document images.

Figure 3. Binarization results for Fig. 1

III. TEXT BOUNDARY LINES

In document page, we call the area of the document except page header and footer main body. In the main body area, the first text line's upper boundary line is named upper text boundary line, the last text line's lower boundary line named the lower text boundary line. In this step, an efficient and robust boundary lines method for the warped document images is proposed. We use line segments to estimate their upper and lower text boundary lines. There steps should be taken as follows:

a) Line segments are used to fit the curve named text upper and lower text boundary lines.

b) Analysis of inflexion is used to improve the result.

c) Curvature based smooth treatment is applied to get the final curved boundary lines.

These steps will be explained in the following parts.

A. Text Boundary Lines Estimation

Because most of Chinese characters except some simple ones in the same line always have the same height, a line segment which is not longer than the character's width can be used as a character's upper baseline and lower baselines which delimit the body of the character. In the unwrapped documents, different from English characters if Chinese characters are in the same line, their baselines are in the same line too, as shown in Fig. 4. These line segments can be connected as text line's boundary line.

Figure 4. Example of charactors' baselines

The line segments could be acquired from the local vertical projection. The length of the segment lines is detected by the width of the characters, which can be detected from the unwrapped area. The length of the character's boundary segment line is 2/3 of the character's width.

Definition_1. Line segment l_2 is connectible only its height is between its left and right adjacent line segments l_1 and l_3.

Fig. 5 gives some examples to illustrate Definition 1. If line segment l_2's height is between l_1's and l_3's height in Fig. 5(a), there is a inequality $H(l_1)<H(l_2)<H(l_3)$, where $H(li)$ is the height of l_i, as the same, $H(l_3)<H(l_2)<H(l_1)$ in Fig. 3(b). While in Fig. 5(c), $H(l_2)$ is not between $H(l_1)$ and $H(l_3)$. Therefore, l_2 in Fig. 5(a) and Fig. 5(b) is connectible, while in Fig. 5(c) is not.

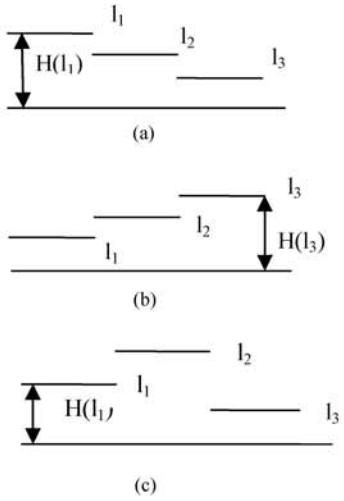

Figure 5. The connectible of line segment: (a) l_2 is connectible; (b) l_2 is connectible; (c) l_2 is not connectible

If the line segment could not be connected, the location of l_2 would be adjusted. The whole procedure of adjustment is as follows,

$$D = H(l_3) - H(l_2) \tag{5}$$

$$H(l_2) = H(l_1) + D/2 \tag{6}$$

here, D is the vertical distance between l_1 and l_3. The results of Fig. 5(c) adjustment is showed as Fig. 5(a).

If the line segment is connectible, two line segments are added, one is between l_1 and l_2, and the other is between l_2 and l_3, as shown in Fig. 6.

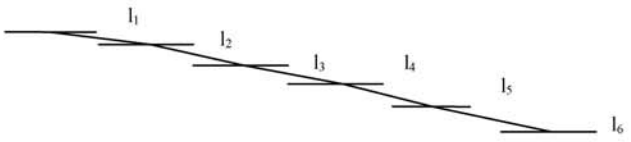

Figure 6. Line segments connected

B. *Inflexion Analysis*

Although some characters are simple and strokes of them are fewer, the segment lines of them may not be accurate. However, the inflexion analysis method which will be proposed can adjust the location of the segment lines' location correctly.

Definition_2. The intersection point of two line segments is an inflexion point only if the clockwise angle between two connected line segments of lower text boundary lines is more than 180 degree, or the clockwise angle of upper text boundary lines is less than 180 degree, as shown in Fig. 7.

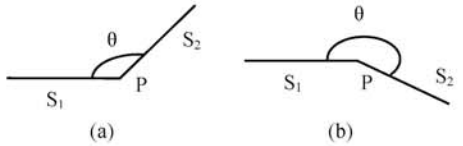

Figure 7. Example of inflexion point: (a) P is an inflexion point in upper boundary lines; (b) P is an inflexion point in lower boundary lines

If there are inflexion points in boundary lines, an inflexion analysis method is used to smooth the boundary lines partly, as shown in Fig. 8.

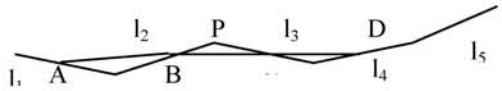

Figure 8. Example of inflexion analysis method

In Fig. 8, P is inflexion point. Therefore, new line segments AB, BC, CD are used to smoothing the boundary lines partly.

C. *Curvature Based Smooth Treatment*

Curvature based smoothing method is used to get the final curved boundary lines. The closer it is to the spine the smaller the curvature of the point is. Curvature of P is shown in Fig. 9.

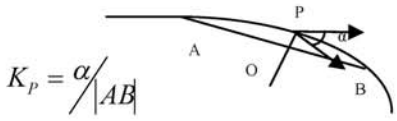

Figure 9. Curvature of the point P

The adjustment of the points' location is in order to make the point having a better curvature, and the final curved boundary lines are showed in Fig. 10.

Figure 10. Example of final result of the text boundary lines

D. *Using Adjacent pages*

Our restoration method is based on the text boundary lines. Therefore, if it fails to get the text boundary lines, the restoration will fail too. In most of the documents in reality, there are text lines existing in the top or bottom area of documents, it makes our text boundary lines could be acquired in most of them. If the top and bottom areas are non-text regions, such as Fig. 11(a), the upper boundary line cannot be acquired. We could make use of the similarity of neighboring pages to restore the document image approximately, and then the follow-up adjacent pages can be used to inspect and further correct the results.

In our method the adjacent pages are the same volume side's adjacent pages. For example, Page N's adjacent pages are Page $N + 2 * i (i = \pm 1, \pm 2, \pm 3, \ldots)$. Because only the same side's pages adjacent have warped similarity. The adjacent pages Page $N + 2 * i (i = -1, -2, -3, \ldots)$ are used to detect the boundary lines approximately, while the pages Page $N + 2 * i (i = 1, 2, 3, \ldots)$ are used to further inspect and correct our approximate results. After that the boundary lines could be acquired more correctly. The final boundary lines of Fig. 11(a) are shown in Fig. 11(b).

(a) (b)

Figure 11. Example of similarity of adjacent pages. (a) Document image to be restored; (b) Adjacent page of Fig. (a); (c) final text boundary lines.

IV. RESTORATION OF DOCUMENT IMAGES

After the above processing, a restoration method based on the upper and lower boundary lines is proposed. Result image of Fig. 1 is shown in Fig. 12.

Figure 12. Results of Fig. 1

As shown in Fig. 10, the text boundary lines have parallel component. Between the two parallel lines is the unwrapped area, while the area between the curved lines is wrapped area. We should adjust the pixels' location of the warped area as same as the unwrapped area. The main idea of our method can be reflected in Fig. 13.

As we known, there are two kinds of geometrical distortion occur when scanning thick bound documents. One is due to projection, and the other is due to lens imaging. The distortion makes information's overlap, reflecting to the obtained images is that some pixels in the warped images are lost. The purpose of restoration is to create and insert these pixels to the obtained images. Therefore, a vertical column-by-column adjustment by nearest interpolation algorithm is performed. Then the lens imaging distortion is restored by bilinear interpolation algorithm.

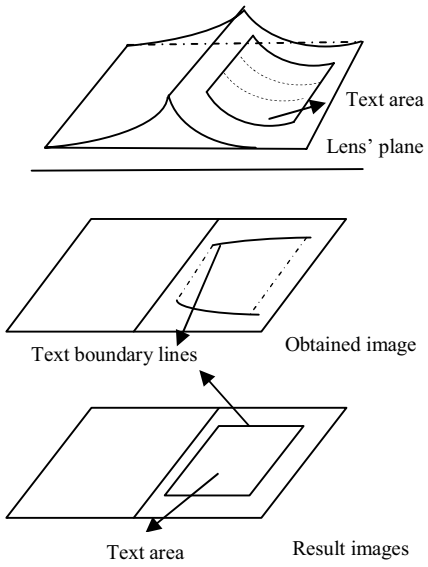

Figure 13. Text boundary lines based restoration

Our restoration algorithm based on upper and lower text boundary lines can be described as follows:

Algorithm of restoration:

1: detect the warped side
   Side=left | right
2: the modified Niblack's algorithm
3: cut off boundary short lines
4: compute boundary segment lines set,
   $L= \{l_1, l_2, ..., l_i, l_{i+1}, ...\}$
5: while ($H(l_{i+1})$ is not between $H(l_{i+2})$*and* $H(l_i)$)
6: $H(l_{i+1}) = H(l_i) + D/2$
7: end while
8: compute connect segment lines' connection point set $m=\{$middle points of segment lines of $L\}$, where $m=\{m_1, m_2, ..., m_i, ...\}$
9: while (l_i and l_{i+1} are connectible)
10: connect (m_i, m_{i+1}) as segment lines s_i
    $S=\{s_1, ..., s_i, ...\}$
11: if (m_{i+1} is a inflexion point)
12: connect (s_i s_{i+1})
13: end if
14: end while
15: while (p is the point of S)
17: adjust K_p
18: end while
19: bilinear and nearest interpolation algorithm.
20: output the result .

V. Experimental Results and Discussions

A. Preprocessing of the documents

The preprocessing methods have been explained in section II. In the warped side detection method, $T=68$ in our experiments, and $k=0.05$ can bring better results than 0.1 in the modified Niblack's algorithm. In section III, when the text boundary lines can not be acquired from target document, its six adjacent pages are used for the text boundary lines' acquirement.

B. Experimental Results

In order to evaluate our method, the OCR testing as a measure of success is carried out both on the original and the result images. The proposed approach has been tested using hundreds of scanned wrapped documents with resolution 150, 200, and 300 dpi. We perform the OCR test using Chinese commercial OCR products, named Shocr6.0 and Hanwang 6.0. The OCR average character precision and recall [13] defined as below and the results are shown in Table I.

$$\text{Character Precision} = \frac{\text{Number of characters correctly detected by OCR}}{\text{Total number of characters detected by OCR}}$$

$$\text{Character Recall} = \frac{\text{Number of characters correctly detected by OCR}}{\text{Total number of characters in the document}}$$

Four hundreds of Chinese documents images are scanned in our database with resolution 150, 200, and 300 dpi. The Shocr6.0 gives both the original images' and the result images' recognition results. Experiments are implemented on a PC with 2.8 GHz CPU and 512 MB memory. The method is programmed by C++ language.

Table I shows the average character recall and the average character precision for the original and result images. The original images are the original scanned document images, and the result images are the restoration images of the original images.

TABLE I. THE OCR RESULTS ON AVERAGE CHARACTER RECALL AND PECISION

Resolution(Num)	Average Character Recall (%)				Average Character Precision (%)			
	Original images		*Result images*		*Original images*		*Result images*	
OCR Product	Shocr6.0	Hanwang6.0	Shocr6.0	Hanwang6.0	Shocr6.0	Hanwang6.0	Shocr6.0	Hanwang6.0
150dpi(100)	57.19	60.22	69.11	72.35	87.12	88.25	95.57	95.80
200dpi(150)	60.78	62.42	75.67	75.72	88.04	89.31	96.45	96.83
300dpi(150)	61.12	62.85	75.45	75.97	88.98	90.09	96.85	97.12

As shown in Table I, the OCR average character recall and precision results are improved by our restored method. For example, the hundred of documents with 150dpi resolution's average character recall are improved from original images' 57.19% to result images' 69.11% for Shocr6.0 OCR product, and the average character precision is improved from original images' 87.12% to result images' 95.57% for Shocr6.0 OCR product. The improvements of them are summarized in Table II.

TABLE II. IMPROVEMENT ON AVERAGE CHARACTER RECALL AND PRECISION

Resolution(Num)	OCR Result			
	Character Recall (%)		*Character Precision (%)*	
OCR Product	Shocr6.0	Hanwang6.0	Shocr6.0	Hanwang6.0
150dpi(100)	11.92	12.13	8.45	7.55
200dpi(150)	14.89	13.30	8.41	7.52
300dpi(150)	14.33	13.12	7.87	6.93

The average character recall is improved by 11.92% to 14.89%, and the average character precision is improved by 7.87% to 8.45% for Shocr6.0 OCR product. From Table I, it can be found that the original images' average character recall is just 57.19%-61.12%, since more than 65% of the words in the shadow can not be recognized by OCR products. After the shadow' removal, the average character recall is improved to 64.06%-69.83%, our restoration method makes it improved to 69.11%-75.45%. Since some words in the shadow are so warped that are even difficult to be recognized by human eyes, the average character recall will be higher while the documents are not warped so much.

In our method, the binarization method to remove shadow makes some warped words' pixels lost that can influence the final recognition rate. The other limitation of our method is that if there are more than five continuous pages whose text boundary lines can't be acquired in the documents, our method may fail. However, this happened less than one percent in reality, and the experiments' results prove our method's success.

VI. CONCLUSIONS

In this paper, we have presented a novel restoration approach which is based on text boundary lines for warped Chinese document images. Our method can remove the shade successfully, and adjust the warped area based on the text boundary lines which are robust even the lines at the top or the bottom of document image are short text lines. Although our method based on the 2D image processing, the images with non-text regions can be restored too. Moreover, our method requires neither page boundaries nor all text lines' contents, nor special setups. The results of our experiments show the effectiveness of the method.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (NSFC 60675025, 60975050) and the National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247, and No.2007AA11Z224), Shenzhen Bureau of Science Technology and Information.

REFERENCES

[1] T. Kanungo, R.M. Haralick, and I. Phillips, "Nonlinear Global and Local Document Degradation Models," Int'l J. Imaging System and Technology, pp. 220-230, Oct. 1994.

[2] T. Wada, H. Ukida, and T. Matsuyama, "Shape from Shading with Interreflections under a Proximal Light Source: Distortion-Free Copying of an Unfolded Book," Int'l J. Computer Vision, vol. 24,no. 2, pp. 125-135, 1997.

[3] M. Pilu, "Undoing Page Curl Distortion Using Applicable Surfaces," Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 67-72, Dec. 2001

[4] M.S. Brown and W.B. Seales, "Image Restoration of Arbitrarily Warped Documents," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 10, pp. 1295-1306, 2004

[5] A. Yamashita, A. Kawarago, T. Kaneko, and K. T. Miura.Shape reconstruction and image restoration for non-flat surface of document with a stereo vision system. IEEE International Conference on Pattern Recognition, 2004.

[6] C.L. Tan, Z. Zhang, L. Zhang, and T. Tao, "Restoration Warped Document Images through 3D Shape Modeling," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 2, pp. 195-208, 2006.

[7] Z. Zhang and C.L. Tan, "Correcting Document Image Warping Based on Regression of Curved Text Lines," Proc. Int'l Conf. Document Analysis and Recognition, pp. 589-593, Aug. 2003.

[8] B. Gatos, I. Pratikakis, K. Ntirogiannis, "Segmentation Based Recovery of Arbitrarily Warped Document Images". Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on Volume 2, 23-26 Sept. 2007 Page(s):989 – 993

[9] A.Ulges,C.H. Lampert & T.M. Breuel,"Document image dewarping using robust estimation of curled text lines ", Proc.ICDAR'05, 2005. pp.1001-1005

[10] X. Zhong, X. Li , Z. Tang, "Correction of Images Scanned form Books" Journal Of Software 2002, Vol.13, No.11 (in Chinese)

[11] W. Niblack. "An introduction to Image Processing ",Prentice-Hall, Englewood Cliff,NJ,pp.115-116,1986

[12] Hong Liu, Qi wu, Hongbin Zha, Xueping Liu, "Skew Detection for Complex Document Images Using Robust Borderlines in Both Text and Non-text Regions", Pattern Recognition Letters, 2008.8.

[13] M. Junker, R. Hoch, and A. Dengle, "On the Evaluation of Document Analysis Components by Recall, Precision and Accuracy," Proc. Int'l Conf. Document Analysis and Recognition, pp. 713-716,1999.