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

A Method to Restore Chinese Warped Document Images Based on Binding Characters and Building Curved Lines

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

Ye Lu
Key Laboratory of Machine Perception and Intelligence
Key Laboratory of Integrated Micro-system
Shenzhen Graduate School, Peking University, China
luye@cis.pku.edu.cn

***Abstract*—With rapid development of information technology, more and more document images are made by scanners. But new problem comes out that many of document images from thick books are warped. It is quite inconvenient for further process on computer. This paper introduces an integrative algorithm on restoring Chinese document images, which is a new filed and few researchers have worked on this subject yet. The complicated structure of Chinese "block words" makes the problem more difficult. To solve this, a restoring method which is based on binding characters iteratively and building curved lines using parallel lines method is introduced. In the phase of fitting, SVR is adopted instead of other parameter methods. An idea of collaboration is also recommended to guarantee the quality of the final results. Correction rate of 94% for experiment of 300 document images proves this method works out very well.**

Keywords—warped document, document image, restoring

I. INTRODUCTION

Modern world is more and more digitized. And large numbers of books in paper could be transformed to electronic form by scanners. In this way, old paper books have their access to this fast moving world. These electronic documents greatly facilitate people's life while cause new problems. As is known to all, computers could not identify words in a document image directly. Therefore, for further process, document images should be turned to other formats that could be used by computers easily. Such mature methods exist if the document images are not warped. Even for skew ones, some methods have been proved effective. But for warped document images which are made by scanners, the problem becomes nonlinear and more difficult. And when the books are quite thick, the situation is worse.

Nowadays, researchers have been working on restoring warped document images. They have 2 main barriers, shading and curved text lines. Various methods could be classified into three categories. Some are stereo or other 3D measurements [1][2][3]. These methods have good precision while need advanced 3D equipments which are not easily accessible. Another kind of methods uses shape-from-shading idea, which needs whole view and solid geometry knowledge [4][5][6]. Similar perspective method could only process camera made document images [7][8].

Model fitting is a method which is based on segmentation. It estimates the warp by fitting some elastic curve or surface model to the text line. This method needs neither calibration, known parameters, nor special devices. And both camera and scanner made document images could use it. Therefore, model fitting is one of the most popular methods in the filed of restoring.

Zheng Zhang and Chew Lim Tan divide the whole document into 2 parts [9]. Model of clean area is a straight line and model of shade area is a quadratic curve. Hironori Ezaki, Seiichi Uchida, Akira Asano, and Hiroaki Sakoe propose a global optimization method [10]. Model is defined as a set of cubic splines and the splines are optimized globally. B. Gatos, I. Pratikakis and K. Ntirogiannis also use similar model fitting but the method firstly drafts de-warped binary image estimation and then recovers the warped image using the transformation factors from the first step [11].

This paper presents a model fitting method which mainly focuses on Chinese characters restoration. It is a new filed of restoring warped document images and few researchers have worked on this subject yet. Due to the structure characteristics of Chinese characters, the problem becomes more difficult and the restoration method needs some new special steps.

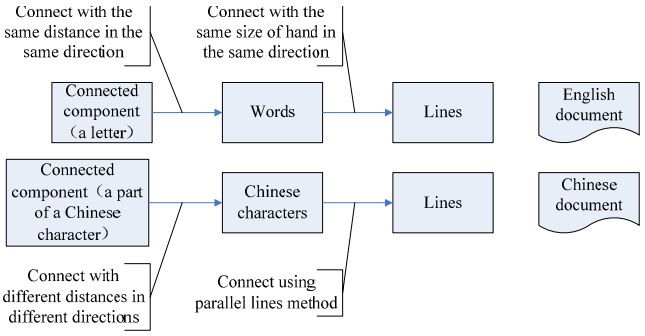


Figure 1. Comparison of processes for building lines in Chinese and English document

Main differences between Chinese and western documents are structure characteristics and shape features of characters.

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

Methods of restoring Chinese and western document images will vary considerably. Fig.1 shows the comparison between English and Chinese document restoring steps.

After segmentation of connected components, the method needs to bind related components to a single Chinese character, and then build related characters to a single curved line. Iterative binding of characters and parallel lines method are separately used here. SVR (Support vector regression) is used during the fitting phase. After a special mistake correction step, idea of continuous restoration plays a role in the last step of this paper.

Collaboration is more and more important in the real world of people. This paper introduces this idea to the world of computer. Each part not only plays its own role during the whole flow but also collaborates with each other to produce better results.

In the following sections, Section II gives a brief framework of the whole algorithm. Section III introduces the isolating model of characters, which includes 2 parts. Part A is about binarization and finding connected components and part B is about how to bind characters. The way to isolate model of curved lines is introduced in Section IV. One part in it is building curved lines and special mistake correction and the other part is SVR fitting. Straightening warped text line by continuous restoring is in Section V. At last, experiments and discussions are presented in Section VI and the whole paper is concluded in Section VII.

II. FRAMEWORK

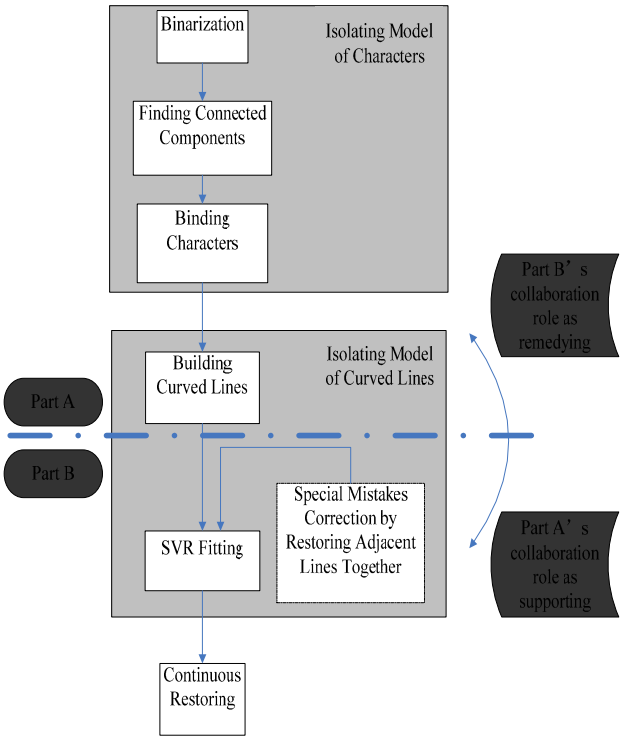


Figure 2. Framework

For document images from scanners, not only curved text lines but also shading should be processed. For thick books, shading is even worse and harder to solve.

Framework of this paper is shown in Fig.2. There are three main phases, isolating model of characters, isolating model of curved lines and continuous restoring. In the first phase, method of adaptive degraded document image binarization [12] is used instead of Niblack's [9]. After finding connected components, iterative binding of characters and "parallel lines" method which is to build curved lines are adopted for Chinese characters' own characteristics. A special mistakes correction step by restoring adjacent lines together is also added. This paper does not divide the whole document into two parts[9]. It does not build models for the two separately but uses SVR fitting. As SVR is a kind of non-parameter method, it just needs to fit each curved line once and does not need to divide. Therefore, it is more reasonable and has better universality. Continuous restoring takes its last role at the end of the whole processes.

In another point of view, the whole framework could be divided into 2 parts, Part A and B, which is separated by the broken line shown in Fig.2. Part A and B both play their collaboration role to each other. Every module in Part A supports the later part for its final restoring. And at the same time, every module in Part B could help to remedy some mistakes which were caused by Part A before. By this way of collaboration, the validity of the method could be guaranteed.

III. ISOLATING MODEL OF CHRACTERS

A. *Binatization and finding connected components*

Adaptive degraded documents image binarization[12] is adopted in this paper. There are three steps, rough estimation of foreground regions, which is the same as Niblack's, background surface estimation and final thresholding.

4-Neighbors Connected Components Labeling Algorithm is used to find connected components in document images [13]. Some too little or too strange (for example, if the result of width divided by height of a component is too large or too small) components should be ignored, because they might be some noises or useless punctuations. And the next step of binding characters is done on the basis of it.

B. *Binding characters*

Structure of Chinese characters is quite different from that of English words. A simple example is "Peking University" and in Chinese it is "北京大学" which is shown in Fig.3. For English words, the definition is to bind characters into a single word. And each English letter is made up of only one connected component (Except "i" and "j". But the dots in these 2 letters could be simply ignored.). Therefore, when binding English words, all that is needed is to connect every letter (shown in blue rectangle) in a word horizontally by the same distance S_1. The former method is to cluster between S_1 (the distance between 2 adjacent letters) and S_2 (the distance between 2 adjacent words), and then uses S_1 to do the binding. The binding direction is shown as red lines.

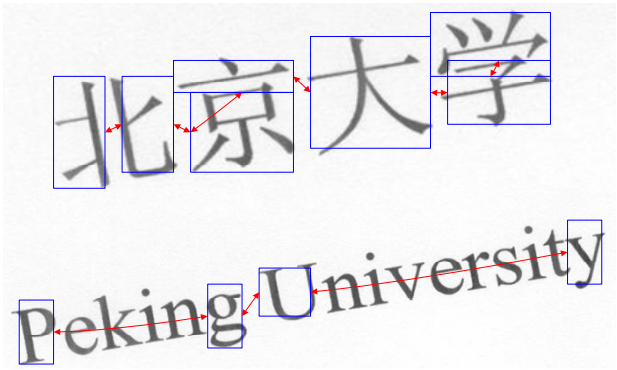


Figure 3. Different structures of Chinese and English

Different from English words, there are a variety of structures for Chinese characters. Two components may be top and bottom. Another two components may be left and right. Or a component may surround another one. The distances between these connected components are different, too. Therefore, this problem becomes complicated.

To solve this, the following definitions and rules are made:

Center_h: horizontal center
Center_v: vertical center
Boundary_hl: left horizontal boundary
Boundary_hr: right horizontal boundary
Boundary_vt: top vertical boundary
Boundary_vb: bottom vertical boundary

$$\begin{aligned} &|Center\_h_1 - Center\_h_2| < d_1 \\ &|Boundary\_hl_2 - Center\_h_1| < d_2 \\ &|Boundary\_hl_1 - Bundary\_hr_2| < d_3 \end{aligned} \quad (1)$$

There are other rules that are symmetrical to (1). And in vertical direction, formulas are similar. By iteratively binding connected components together according to these rules, all the Chinese characters could be got. Although there are some binding mistakes, it will not influence the restoring result under the framework of collaboration, which will be further explained in Section IV and Section V.

Details of iterative binding are shown in Fig.4.

```
Input: Region *R
Output: Word *W
/*Iterative binding*/
1. i=0
2. while(IsEmpty(R)==false)
3.     i++
4.     for j:=0 to sizeof(R) do begin
5.         if(IsNotUsed(R[j])==true)
6.             if(IsEmpty(W[i])==true)
7.                 Initialize W[i]
8.                 Insert R[j] into W[i]
9.                 Delete R[j]
10.            end
11.            else
12.                for all R[k] in W[i]
13.                    if(Rules(R[j],R[k])==true)
14.                        Update W[i]
15.                        Insert R[j] into W[i]
16.                        Delete R[j]
17.                        break;
18.                end
19.            end
20.        end
21.    end
22. end
23. end
```

Figure 4. Iterative binding

IV. ISOLATING MODEL OF CURVED LINES

A. *Building curved lines and special mistake correction*

Reference [9] deals with building curved lines using modified "box-hands" method. When it comes to Chinese characters, new problem is that we do not have a way to find the angle of hands due to structure of characteristics. Therefore, a new "Parallel lines" method is worked out, as shown in Fig.5. Firstly, rough lines segmentation is given by horizontal projecting of the more straight side. For left pages it is on the left and for right pages it is on the right. This step is to find the first character in each line on the more straight side. Then the next word is found on one of the parallel lines which is nearest to the previous word. If this image is from a right page of a book, the words are found from right to left, so our finding scope is A. Because 2 adjacent words could not be too far away, a rectangle limit is also given, shown as broken line. This method could also save a lot of time cost because it cuts the step of clustering distance S_1 and distance S_2.

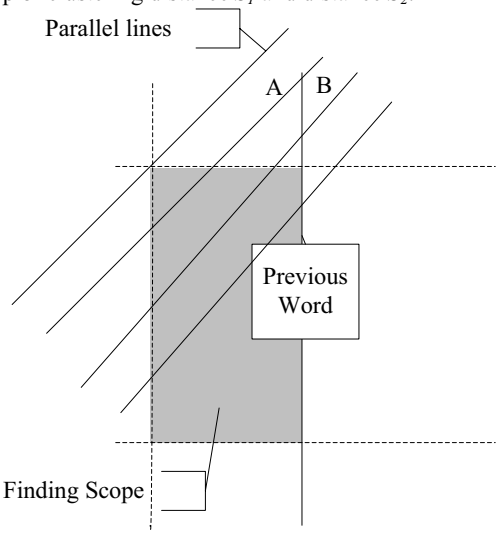


Figure 5. Parallel lines method

Some special lines should be paid attention to, like some ones which start too late or are too short. They should be given up because they will destroy the quality of the following SVR fitting or continuous restoring steps. Therefore, this method uses the idea of binding adjacent lines to the given up line. And these lines are restored as a group according to the data

criteria of adjacent lines. In this way, the difficulty of special lines mistakes could be overcome. And this is the first place where Part B remedies Part A's faults to collaborate on giving accurate curved lines.

B. SVR fitting

After the former steps, all the characters in each line could be marked. The next task is to finish a curved line fitting according to the data from these characters. To explain more, it means that we need to find how the curved line of Chinese characters is like.

Data resource should be considered first. English words have flat bottom boundary which could be used as data resource. The bottom boundary center of each word is often adopted. While Chinese characters have both flat bottom and top boundaries, which are called "block words", here the center of each block character is adopt.

SVR fitting is used instead of parametric fitting like straight line model or quadratic curve model. Reference [9] divided the whole document into 2 parts and built 2 kinds of models. In this way, there must be some incontinuity at the dividing place. Non-parametric SVR fitting method builds a uniform model for each curved line and does not need 2 parts any more. Therefore, it is a more reasonable method and has better universality.

Chih-Chung Chang and Chih-Jen Lin's LIBSVR tool is used in this paper [14]. Epsilon-SVR is chosen and for the kernel function:

$$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j) \tag{2}$$

Basic kernel is chosen as radial basis function (RBF):

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|) \tag{3}$$

Fig. 6 shows the result of SVR fitting. In this way, all the curved lines of a document image could be got from data resource which was introduced.

Principle of SVR fitting decides the result will not be affected by several mistaken data. This is the second place where Part B remedies Part A's faults to collaborate on giving non-parametric models and not influenced by binding characters' mistakes.

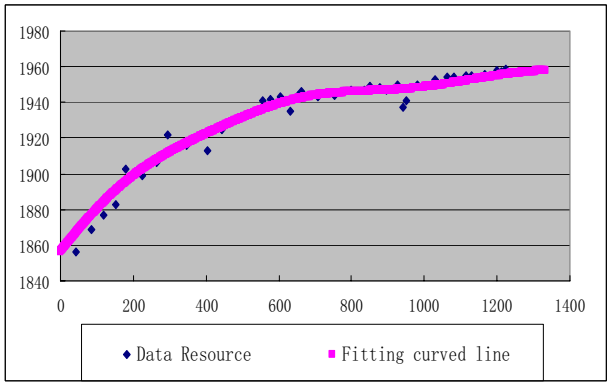


Figure 6. SVR fitting

V. STRAIGHTENING WARPED TEXT LINE BY CONTINUOUS RESTORING

After all the above processing steps, the fitting curved character lines are got. All left to do is to straighten them. Reference [9] straightened by unit of "words", which were the "bounding boxes". But as talked about before, Chinese characters have more complicated structures. If the same method were used, some connected components which were not included in the "bounding boxes" or were not correctly included would cause discontinuous restoring and have a bad performance on final results.

To solve this problem, a continuous restoring method for straightening warped text lines is designed. As shown in Fig.7, this problem has such a modeling. Several curved lines given (shown as the red thick lines), every dot at the vertical line shown in the most right of the whole page needs another curved line (shown as the blue thin lines) which passes through itself and is a fitting of the two nearest adjacent lines.

In Formula (4), y is in the vertical direction. *line_result(y)* shows the fitting curved line of the two nearest adjacent lines. *line1 (y)* and *line2 (y)* represent the lines above and below. *most_right* means the index is at the most right of the lines. w_1 and w_2 are 2 relevant weights.

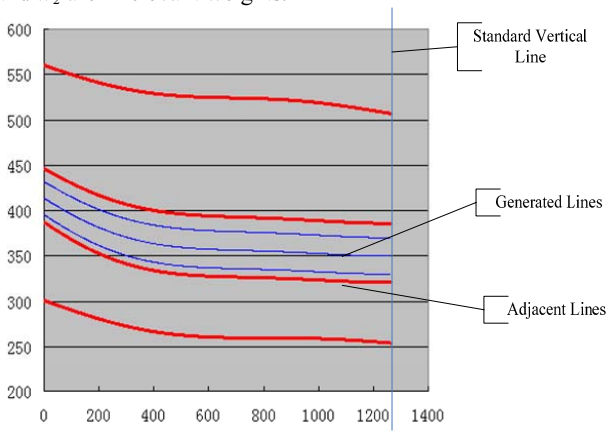


Figure 7. Continuous restoring model

$$line\_result(y) = \frac{line_1(y) \times w_1 + line_2(y) \times w_2}{w_1 + w_2}$$
$$w_1 = line_2(most\_right) - line\_result(most\_right) \tag{4}$$
$$w_2 = line\_result(most\_right) - line_1(most\_right)$$

Under the rule of Formula (4), a group of continuous curved lines could be got. And each pot on this document image could transform to a new place according to the same method [9] has suggested.

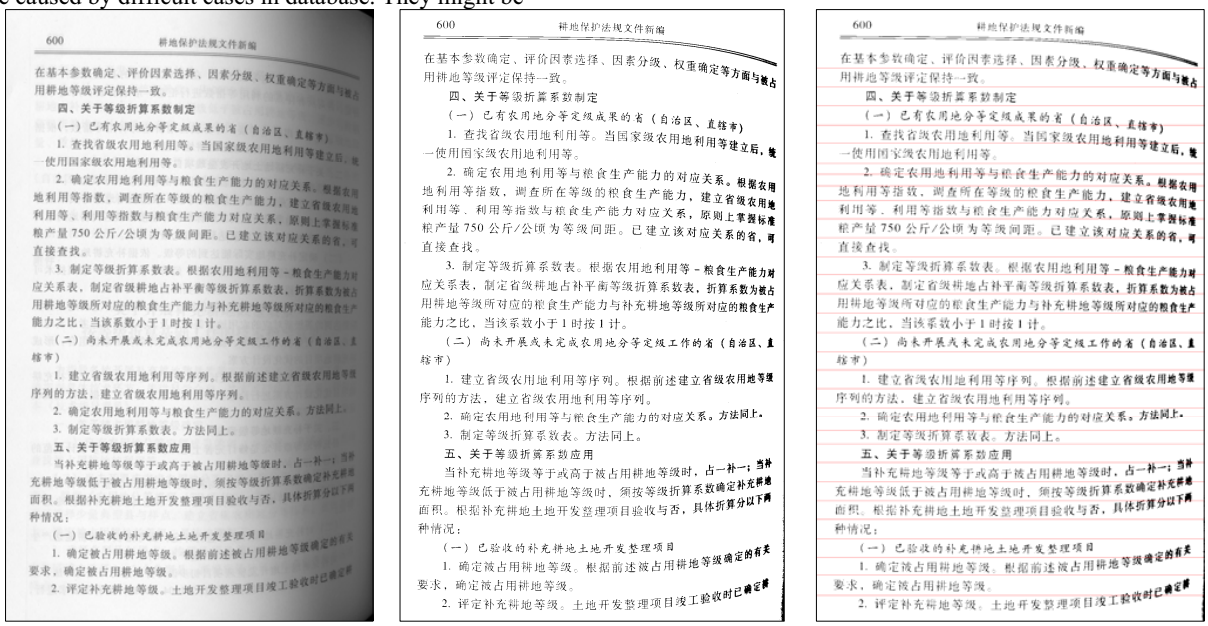
Continuous restoring will help to envelop some mistakes caused by former steps. This is the last place where Part B remedies Part A's faults to collaborate on using continuous restoring idea.

VI. EXPERIMENTS AND DISCUSSIONS

The method is implemented on a database from a thick book which has 915 pages. The database is made up of 300 random pages chosen from this book. A final correction rate of 94% is got.

From Table I, it could be seen that the 2% failing pages were caused by difficult cases in database. They might be document images whose text lines have very large slopes or those which have severe noises.

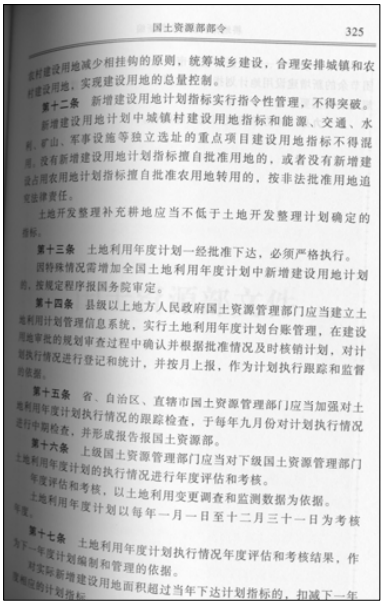
Another 2.3% failing pages happened during stage of building curved lines and the rest 1.7% happened during continuous restoring stage. None of the failure cases was caused by isolating model of characters or continuous SVR fitting.



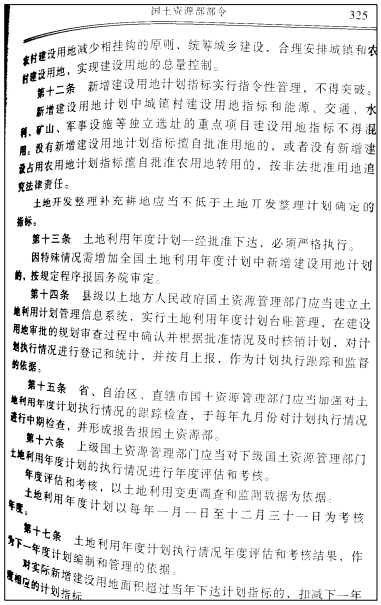
(a) Original page (b) Binary result (c) Rough lines segmentation

(d) Binding characters and building curved lines (e)Final result

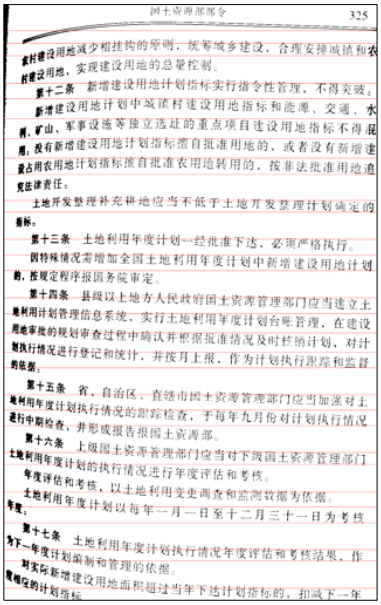
Figure 8. Experiment results of left page



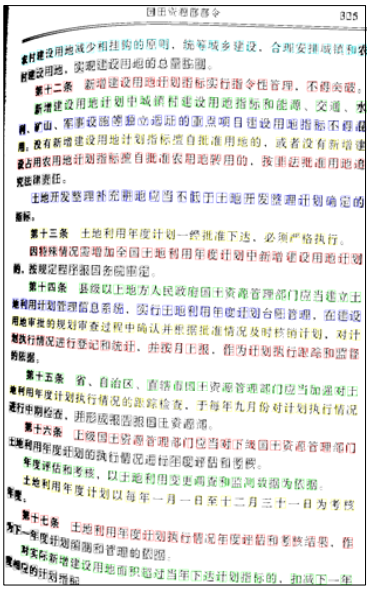
(a) Original page



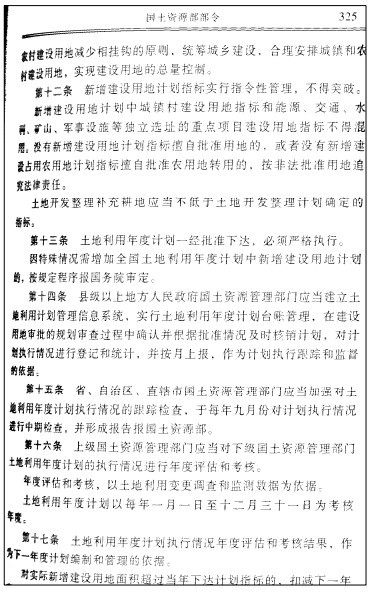
(b) Binary result



(c) Rough lines segmentation



(d) Binding characters and building curved lines



(e)Final result

Figure 9. Experiment results of right page

Results show that the method is robust during each step. Not a single failure was caused by isolating model of characters, which means that the collaboration idea of remedying Part A's results by Part B is quite effective. SVR fitting also did well in experiments which proved that it is better than other parametric methods. The 1.7% failure cases due to continuous restoring show that this step caused some new kind of flaw to the method. But considering its advantage on improving the efficiency of the former stage, especially that of Part A, it is still worth adopting.

Fig.8 and Fig.9 are some of the document images from the experiment. Fig.8 gives a handling process of a left page and Fig.9 gives that of a right one from the database. The original pages, binary results, rough lines segmentation, binding characters and building curved lines and the final result document images are given respectively.

In (a) original page, it could be seen that both shading and warped lines are severe. After adaptive degraded documents image binarization, in (b) binary result it is clear that shading has already been discarded but warped lines were the same. In

(c) rough lines segmentation, each curved line is separated by red straight lines. It is surely not very accurate on the side which is nearer to the spine of the book, but it is enough for later processing because only the first few characters in each line are needed from the former step to build the curved lines. In (d) binding characters and building curved lines, each curved line is marked with a single color. Some lines are not marked, which means they are "special lines" which need a step of restoring adjacent lines together. The final result is shown in (e) and at this time both the problems of shading and warped lines have been solved.

TABLE I. EXPERIMENT RESULTS

	Pages/p	Rate/%
Mistakes Analysis		
Difficult Cases in Database	6	2
Isolating Model of Characters	0	0
Building Curved Lines	7	2.3
SVR Fiitting	0	0
Continuous Restoring	5	1.7
	Total	Correction Rate
Left Pages	150	95.3
Right Pages	150	92.7
Sum	300	94

Although there might be some mistakes during each step, they will not affect the final results very much. It benefits from the idea of collaboration.

While this method still has some limitations. Since it is based on continuous restoring, there will not be any mistake of restoring single box of word, but the emphasis on the continuity makes it not very accurate near top and bottom boundaries of the page.

VII. CONCLUSIONS

This paper introduces a new method to restore warped documents images, especially for Chinese ones. After binarization and finding connected component, a key step is to iteratively bind characters in different directions and build curved lines using parallel lines method. This step is quite important because it is the main difference between flows for Chinese and English documents. Special mistakes correction by restoring adjacent lines together and SVR fitting are also adopted for better restoring quality. At last, continuous restoring is used.

An idea of collaboration is also adopted. The whole flow is divided into two parts, Part A and B. Part B is the important restoring phase so Part A plays a supporting role to do some preparing work for B. At the same time, Part B plays a remedying role, which makes sure that some mistakes caused by Part A will not affect the final result very much.

Experiments of 300 document images with a correction rate of 94% show that this method does well in the field of Chinese document images restoration. Our future plan is to apply the method on complicated documents with different type sizes or on those which include pictures and tables.

ACKNOWLEDGMENT

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

REFERENCES

[1] Michael S. Brown, W. Brent Seales, "Image Restoration of Arbitrarily Warped Documents", IEEE Transactions on pattern analysis and machine intelligence, Vol.26, No.10, 2004, pp.1295-1306.

[2] Mingxuan Sun, Ruigang Yang, LinYun, George Landon, Brent Seales, Michael S. Brown, "Geometric and Photometric Restoration of Distorted Documents", Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV), Vol.2, 2005, pp.1117-1123.

[3] Michael S. Brown, Mingxuan Sun, Ruigang Yang, Lin Yun, W. Brent Seales, "Restoring 2D Content from Distorted Documents", IEEE Transactions on pattern analysis and machine intelligence, Vol.29, No.11, 2007, pp.1904-1916.

[4] Li Zhang, Chew Lim Tan, "Restoring Warped Document Images using Shape-from-Shading and Surface Interpolation", Proceedings of the 18th International Conference on Pattern Recognition (ICPR), 2006, pp. 1932-1936.

[5] Zheng Zhang, Chew Lim Tan, Liying Fan, "Estimation of 3D Shape of Warped Document Surface for Image Restoration", Proceedings of the 17th International Conference on Pattern Recognition (ICPR), Vol.1, 2004, pp.486-489.

[6] Yau-Chat Tsoi, Michael S. Brown, "Geometric and Shading Correction for Images of Printed Materials. A Unified Approach Using Boundary", Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol.1, 2004, pp.I-240-I-246.

[7] P.Kakumanu, N.Bourbakis, J.Black, S.panchanathan, "Document Image Dewarping based on Line Estimation for Visually Impaired", Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2006, pp.625-613.

[8] Camille Monnier, Vitaly Ablavsky, Steve Holden, Magnús Snorrason, "Sequential Correction of Perspective Warp in Camera-based Documents T", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR), Vol.1, 2005, pp.394-398.

[9] Zheng Zhang, Chew Lim Tan, "Correcting document image warping based on regression of curved text lines", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR), vol.1, 2003, pp.589-593.

[10] Hironori Ezaki, Seiichi Uchida, Akira Asano, Hiroaki Sakoe, "Dewarping of document image by global optimization", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR), Vol.1, 2005, pp.302-306.

[11] B. Gatos, I. Pratikakis, K. Ntirogiannis, "Segmentation Based Recovery of Arbitrarily Warped Document Images", ICDAR, Vol.2, 2007, pp.989-993.

[12] B. Gatos, I. Pratikakis, S.J. Perantonis, "Adaptive degraded document image binarization", Pattern Recognition, 2006, pp.435-440.

[13] L.Di Stefano, A.Bulgarelli, "A simple and efficient connected components labeling algorithm", International Conference on Image Analysis and Processing, 1999, pp.322-327.

[14] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM : a library for support vector machines", Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.