Restoration Method of Distorted Digital Document Image Based on Text Line Detection

Chong Shen[1, a], Lijing Tong[1, b], Jian Zhan[1, c], Zaiyin Zhang[1, d]

[1]Department of Computer Science, North China University of Technology, Beijing, China

[a] 690023772@qq.com, [b] 78021968@qq.com, [c] 100176982@qq.com, [d] 1170085047@qq.com

Keywords: digital document image, distorting, restoration.

Abstract. More and more documents are scanned into digital image, meanwhile, the document scanned into digital image will appear the phenomenon of widespread distortions and shadows. Distortions in a variety of document images have an impact on people's reading comprehension or automated document image processing. In order to solve the problem, this paper uses image segmentation technology to detect text lines for getting lower baseline and upper baseline of the text. With the lower baseline and upper baseline we can adjust the distortion of the document image. So that the corrected image can be obtained. In the image pre-processing, sharpening is a pivotal step.

Introduction

At present, for the distorted digital document image, there are three major correction method as follows:

1) Geometric distortions are corrected by sending horizontal and vertical vanishing points toward infinity in a down-sampled image. Moiré pattern noise is removed using low-pass filters with different sizes independently applied to the background and text region. The contrast of the text in a specular highlighted area is enhanced by locally enlarging the intensity difference between the background and text while the noise is suppressed [1].

2) Warping shape of each text line is acquired by estimating baselines' shape and characters' slant angles after line segmentation. In order to get fluent recovery result, thin-plate splines are exploited whose key points are determined through the result of warping estimation [2].

3) Based on geometry of differential and projection as well as the theory of imaging optics, a robust and fast rectification approach is proposed for the restoration of camera images of planar and curled document [3].

Text Line Detection Based Restoration Method

The proposed method will use four steps to restore any geometric distortion in document image. Step One is Sharpening. Step two is Grizzling. Step three is Binarization. Step four is Image Stretch.

A. Sharpening.

The image sharpening is a contour compensation, protruding edge information to make the image clearer. The goal is to substantially enhance the high frequency components of the original image [4]. The basic algorithm is as follow:

$$g(x,y) = | w(x,y) * f(x,y) | \quad (1)$$

where $f(x, y)$ is the original image, $w(x,y)$ is used to control the sharpness of the image. "*" denotes the convolution operation. Obviously, Eq.1 eliminates the majority of low frequency

components of the original image, and retains the high frequency part. The size of neighborhood window is 3*3. We calculate the image of each point *with* Eq.1 and get the *g(x, y)* value. For example, the mask of Eq.1 is:

$$w(x,y)=\begin{bmatrix}-1 & -1 & -1\\ -1 & 9 & -1\\ -1 & -1 & -1\end{bmatrix} \qquad (2)$$

The result of this method is as follows. Fig.1 is original image. Fig.2 is sharpened image.

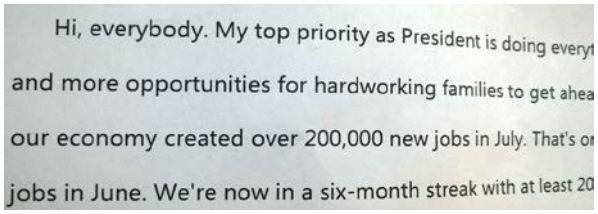

Fig.1 Original image

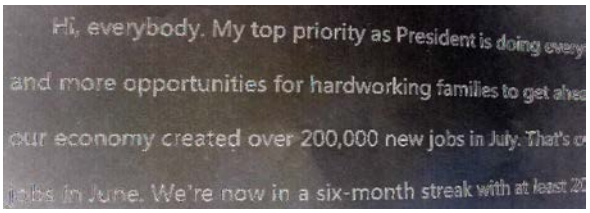

Fig.2 Sharpened image

B. Grizzling.

In order to simplify the information of images and further processing, we use Eq.3 to grizzle the document images [5]. The Eq.3 is:

$$Gary(i,j)=0.299*R(i,j)+0.587*G(i,j)+0.114*B(i,j) \qquad (3)$$

where *R(i,j)* is the red component. *G(i,j)* is the green component. *B(i,j)* is the blue component. With the Eq.4, we can obtain a grizzled image as Fig.3.

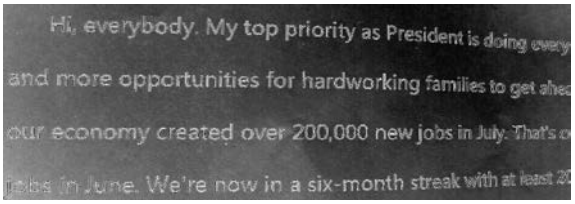

Fig.3 Grizzled image

C. Binarization.

Here a global threshold method is used with global iterative algorithm [6]. Firstly, we chose an initial threshold which is the mean of the biggest gray value and the minimum gray value. With the threshold, image is divided into target and background. Then, we set the mean of the target and the background as the new threshold. With the new threshold, image is divided into target and background, and so on. When the threshold doesn't change, we get the final threshold. With the final threshold, we binarize the image. If gray level is less than the threshold, we set it 0. If gray level is greater than the threshold, we set it 255. Details of global iterative algorithm is as follows:

1) Mean of the biggest gray value $g_{\max}$ and the minimum gray value $g_{\min}$ is the initial threshold:

$$T_K=\frac{g_{\max}+g_{\min}}{2} \qquad (4)$$

2) According to the threshold *T*, the image is divided into target and background. And we calculate the number of pixels of target and background, N_t, N_b. Then we calculate the two parts' mean of gray value respectively, $\overline{g}_t, \overline{g}_b$. Formula is as follows:

$$\overline{g}_t = \frac{\sum g(i,j)}{N_t},\ \overline{g}_b = \frac{\sum g(i,j)}{N_b} \qquad (5)$$

3) According to *Eq.5*, we can calculate the new threshold. Formula is as follow:

$$T_{K+1} = \frac{\overline{g}_t + \overline{g}_b}{2} \qquad (6)$$

4) If $T_K = T_{K+1}$, the final threshold is T_K. If not, do the step 2 again.
The result of this method is as follow. Fig.4 is binarized image.

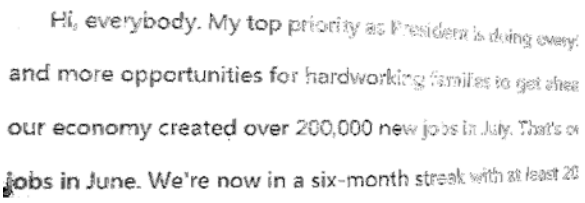

Fig.4 Binarized image

D. Image Stretching.

Firstly, each pixel line is scanned in the binarized image. Then pixel number of black point of each pixel line is calculated, so upper boundary and lower boundary are found. According to the upper boundary, a distortion starting point is found. According to the lower boundary, a distortion bottom point is found. With the distortion starting point, the distortion bottom point and height of word, distortion coefficient is calculated. With the distortion coefficient, the distorted digital document image is restored. Details of the method is as follows:

1) Here the threshold of black point number in the text boundary line is set as *T*, during the pixel line scanning. If pixel number of each line is more than *T*, the pixel line is judged as a boundary.

2) Here the upper-left corner of the image is set as the original point of the coordinate system. For the average pixel height of words is *H*, upper boundary's Y-axis value is y_1, a suitable Y-axis value of standard text line can be calculated as $y_1 + H/2$.

3) The upper boundary and the lower boundary, found in step 1, are scanned from left to right. Then the distorted starting point and the distorted ending point are identified. Supposing the distortion starting point's x-axis value is x_1, the distortion ending point's x-axis value is x_2, the distortion coefficient D is calculated as Eq. 7.

$$D = (y_2 - y_1 + H/2)/(x_2 - x_1) \qquad (7)$$

4) From the distortion starting point to the distortion ending point, we raise each $point(x, y)$ up. The height of each point raised is C. C is calculated as Eq. 8. Finally, the restored digital document image is obtained.

$$C = (x - x_1) * D \qquad (8)$$

Experimental Results and Analysis

Hanwang *OCR6.0* recognition results is used to compare and analyze here. Fig.5 is the original image. Fig.6 is the result image. The result of recognition rate comparison is listed as Table 1.

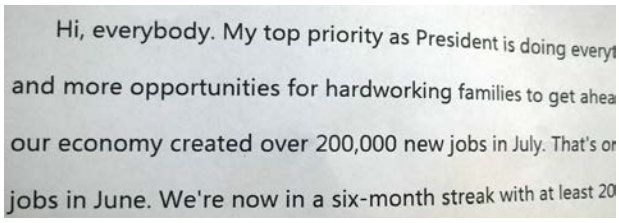

Fig.5 Original image

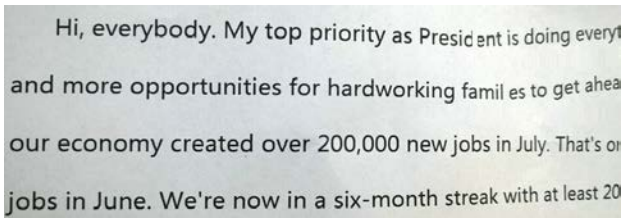

Fig.6 Result image

Table 1 Recognition Rate Comparison

	Number of Letters	Number of Correct Identified Letters	Recognition Rate
Before Restoration	195	148	75.90%
After Restoration	195	177	90.77%

Conclusion

In this paper, a restoration method of distorted digital document image based on text line detection is proposed. The method can correct distorted document image effectively and the OCR recognition rate of the corrected image is significantly higher than the image which has not been corrected.

Acknowledgement

This research is funded by National Natural Science Foundation of China (61371142), the project(14005) and Science activities for university students in 2014 are supported by North China University of Technology.

References

[1] Simon Christian, William, Park In-Kyu: Correcting geometric and photometric distortion of document images on a smartphone, Journal of Electronic Imaging, v 24, n 1, January 1, 2015

[2] Cs Liu, Y Zhang, Bk Wang, Xq Ding: Restoring camera-captured distorted document images, International Journal on Document Analysis and Recognition, November 26, 2014

[3] Xuejing Dai:A novel approach for the restoration of camera images of planar and curled document, Computer Science and Education (ICCSE), 5th International Conference,2010

[4] J.L. Zeng: An Image Sharpening Algorithm Based on Edge Detection, modern electronic technology journals, 2006, p231

[5] D.S Chen, F.F.Song, Q. Zhang: An Adaptive Global Mapping Approach for Color to Gray Image, Conversion Journal of Computer Systems & Applications, 2013, p324

[6] Chi, Z., A two-stage binarization approach for document images, Video and Speech Processing, 2001