A Correction Algorithm for Document Images Based on Edge Contour

Jianhao DING  
*School of Media and Design, Hangzhou Dianzi University, Hangzhou, China*

Zhijie LIN  
*Zhejiang University of Science and Technology, Hangzhou, China*

Lingyun YU  
*School of Media and Design, Hangzhou Dianzi University, Hangzhou, China*

**ABSTRACT:** For digitized document images with black border, this paper presents a simple and rapid correcting method. Firstly, we extract the contours using canny operator and find the inner contour with maximum area. The bounding box of the contour is used to calculate the inclined angle as the skew angle of the document image. According to the skew angle, we rotate the image around the bounding box center to correct it. Lastly we use the average color of the image to fill the skew corners due to correction and crop the black border outside the bounding box. The experiment results indicate that the presented algorithm is exact, efficient, also with robustness.

**KEYWORD:** Algorithm; Document Image; Edge Contour

1 INTRODUCTION

In the field of document image processing, most of the OCR(optical character recognition) system is very sensitive to the skew of document images, the quality of skew correction direct impact on the quality of the image processing. Therefore the researches on skew correction for scanned document images have a great significance in most real-time document image processing. Skew angle should be confirmed quickly and accurately to improve the accuracy of collection and entry for document information. A number of good skew estimation algorithms are available in the literature. However, the time required to estimate the skew angle is still an important issue.

The existing methods of skew angle estimation basically have several categories [1]: the method based on Hough transform, the method based on cross-correlation, the method based on projection, the method based on fourier transform and the method of K-Nearest Neighbor. Literature [2] proposed run-length smoothing algorithm for skew detection and text line position determination in digitized documents, which is based on the cross-correlation between the pixels of vertical lines. Literature [3] used a text line projection to transform, and then uses the hough transform to estimate the skew angle of document images. A few algorithms capable of detecting both skew and orientation for Roman script document images have appeared [4,5]. For documents making heavy use of majuscules or using scripts other than roman (e.g. Pashto, Hindi, Arabic), different decision criteria for orientation detection were recently introduced in [6]. Based on the analysis of texture complexity of document image with homogeneous background, He proposed a skew-detection and skew-adjustment algorithm [7].

Almost all methods outlined above work in a limited angle range, have limited angle accuracy, and need a lot of input data and computation time. The algorithm proposed in this paper is simple, fast and works with any kind of objects, like bend lines (not only straight lines), pictures. For digitized document images with black background such as figure 1 shown, our method does not have to consider the inner complex structure of documents, and only need to correct the document boundary.

Figure1. Some samples of digitized document image with black background
2 CONTOUR DETECTION

A binary image is a two-dimensional binary function whose values are 0 or 1. Binarization is the premise of contour extraction. We use the average gray value of the whole image as the threshold to compute binary image. In this work we have used mathematical morphological operations, i.e., dilation, erosion, opening, and closing to smear the contour to solid black bands and to remove the bumps in the bands. The contour is extracted using canny operator [8] in this paper, which is illustrated by figure 2.

From figure 2, the detecting process of the canny algorithm consists of the following steps. The algorithm used a functional derivative of Gaussian function to deduce the first derivative is the best edge detection operator’s best approximation. He utilized the first order differential operator to calculate each point gradient amplitude and gradient direction to obtain the corresponding gradient amplitude image and gradient direction image. In order to locate the edge accurately, non-maximum suppression for the gradient was processed. Dual-threshold method was used to detect edge points from the candidate and the connection point of the final edge lastly. The figure 3 shows the results of our algorithm in each step.

3 CORRECTION OF DOCUMENT IMAGE

3.1 Skew angle calculation

Since the boundary shape of document content usually appears to be a rectangle, the literature [9] assumes that there exists a minimum rectangle that contains the complete content region of a given document image. Our approach collects a double-layer structure contour from edge Image. The first layer is the contours including the inner contour and the bounding box, while the second layer composed of the contours with hole. We traverse the inner contours to find the contour with maximized area, which contains the entire document region. In figure 4, the red line is the bounding box of the contour indicated in green color.

3.2 Image rotation

The skew angle $\theta$ is obtained by the corner coordinates of the bounding box, as the figure 5 shown. When the scanned image is inclined, in order to obtain the original images, it should be rotated around the image center with a certain angle. The rotated result is translated so as to put the whole image into the display area. The algorithm confirms the rotating center $\left(x_c, y_c\right)$, which is the image center generally according to the bounding box, as the figure 4 shown. The corrected pixel point $\left(x', y'\right)$ is obtained by the formula:
4 EXPERIMENTS

To test the performance of our algorithm on complex document images, we run the rectangle detector on the dataset consists of 100,000 images from document digitalizing enterprise. Figure 6 shows some skew detection results by our algorithm. The top row is the initialized digitalized image using camera.

We use the average color of the image to fill the skew corners due to correction and crop the black border outside the bounding box.

Figure 5. Skew angle of the bounding box

Figure 6. Some samples of our algorithm
The second row is the contour with maximized area extracted using canny operator. The third row shows the bounding box of contour with red lines. The bottom row is the results with correction and cropping the black borders. Our method achieves 4.5% miss rate, which cannot find the bounding box accurately.

Some samples of false bounding box of content region are shown in Figure 7. Most false alarms are generated due to exist larger contour closing to the edge of the document such as the covers.

Figure 7. Part false samples from our approach

5 CONCLUSION

In this paper, an efficient, high precision algorithm is presented for the correction of the skew angle of scanned document images. The proposed method applies to any kind of documents even when they contain images mixed with text. Because our algorithm does not rely on local image features and it applies global image features to drive the rectangular curve close to the boundary of content region, it can detect the skew angles accurately in those complex images processed unsuccessfully by other methods. Our algorithm differs from other skew detection methods in that it does not rely on local image features. Instead, it uses global image features and shape constraint to obtain a strong robustness in detecting skew angles of document images.

ACKNOWLEDGEMENTS

The research work was supported by Foundation of Zhe’jiang Educational Committee under Grant No.Y201224034.

REFERENCES