Rectification of Arabic Text Document Images Linearly and Curved Distorted on Acquisition

1Agus Zainal Arifin, 2Anny Yuniarti, 3Erliyah Nurul Jannah
12Department of Informatics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia
3Universitas Pesantren Tinggi Darul Ulum (UNIPDU), Jombang, Indonesia
1agusza@cs.its.ac.id, 2anny@if.its.ac.id, 3erliyah.nj@gmail.com

Abstract

Linear and curvature distortion often occurs in the text document image acquired by scanner or digital camera. Optical Character Recognition (OCR) software may not recognize the characters well, especially for Arabic text. In this paper, we proposed a new method for rectification of Arabic text document linearly and curved distorted on acquisition. The proposed method is a segmentation-based method which applied for Arabic text that has different characteristic than others. It consists of bounding box reduction, slope estimation and correction, text line correction, and image reconstruction. Estimation of slope involves simple linear regression, and the number of data in the regression will be limited to get better estimation. Our experiment shows that the proposed method able to estimate the slope of the word more accurate than the conventional segmentation-based methods and achieves more number of character recognized by OCR software.

Keywords: linear distortion, curvature distortion, OCR, segmentation-based method, simple linear regression.

1. Introduction

There are many old documents printed in Arabic text. It is very important to maintain such documents, especially the documents of Islamic or Arabic heritage. Capturing and scanning the documents are needed for digitalization and recognition. However, inappropriate scanning may lead geometrical distortion including curvature and linear distortion as shown in Figure 1. Optical Character Recognition (OCR) software will not recognize the characters well. Several methods have been proposed for rectifying such distortion [1,2,3].

In fact, SEG [4] is the most robust method among the other method, but it could not achieve a good performance if applied for Arabic text document. Arabic text has different character from others. Latin text tends to have similar height of letters, while Arabic text tends to vary its height. In this case, linear and curvature distortion which occurs in the Arabic text document image may lead to unsuccessful recognition by OCR. In this paper, we proposed a new method for rectification of Arabic text document linearly and curved distorted on acquisition.

2. Method and Implementation

Rectification includes bounding box reduction, slope estimation and correction, text line correction, and image reconstruction [4]. Slope estimation involves simple linear regression analysis. In this case, Arabic document needs more analysis in estimation of word slope than Latin text document.

![Sample of geometrically distorted document images](image)

(a) (b)

Figure 1. Sample of geometrically distorted document images; (a) Linearly distorted document images; (b) Curvature distorted document images.

The first step of word slope estimation is detecting x-axis and y-axis of uppermost and lowermost black pixel. Next step is taking an average from both x-axis and y-axis data for each word. The x-axis within the
range \([\mu_x - t, \mu_x + t]\) will be the independence variable, and y-axis within the range \([\mu_y - t, \mu_y + t]\) will be dependence variable on simple linear regression. Threshold \(t\) is an empirical value with range between minimum and maximum of axis data value. Each image has different \(t\). By this linear regression, slope of every word will be obtained. Correction of word slope can be done by rotating the word based on their slope value. Every word will be rotated to horizontal line.

3. Experiments

The dataset used in this paper consist of 10 Arabic text document images. Those images have different distortion (curvature or linear distortion) and text fonts. We do the experiments to get best value for threshold \(t\). The best value of \(t\) is occurred when we get the most character recognized by OCR software with that value of \(t\). Figure 2 is the example of experiment results.

Our experiments result show that 7 images have increased number of character recognized by OCR software, 1 image has decreasing, and 2 images have no increasing or decreasing. \(O\) is the number of character recognized from original images (without rectification). \(R\) is number of correct character recognized from result images (after rectification). And \(T\) is total character manually calculated from every sample image.

4. Discussion

Limited number of characters recognized by OCR is occurred because of improper parameter selection. Improper parameter for bounding box reduction [4] will affect the detection of text line. If the text line detection is not accurate, then the rectification process failed. Determination of threshold \(t\) also can influence the number of character that can be recognized by OCR. Each image has different \(t\). However, images with same text fonts usually have same \(t\) due to similar characteristic, such as character height proportion.

Difficult recognition may be happened on the document contained complex character as shown in Figure 2 (c) and (d). Although rectified the image had been rectified, the OCR software still can not recognize the character due to its complexity.

We used OCR for evaluation, since it seems more objective than manual evaluation. We measure the ratio between the number of character that correctly recognized by OCR software and total character in the text document image. Our proposed method has better performance than segmentation based rectification (SEG) in providing input image for OCR. If SEG applied to Arabic documents image, no character can be recognized by OCR or 0% accuracy. But our proposed method can perform better. The character that can be correctly recognized by OCR software is 22.42% in average (for 10 images).

![Figure 2](image)

Figure 2. Figure a,c,e,g are original image (before rectification); Figure b,d,f,h are result image (after rectification).

5. Conclusion

This paper proposed a new method for rectification of Arabic text document linearly and curved distorted on acquisition. The rectification process consisted of bounding box reduction, slope estimation and correction, text line correction, and image reconstruction. We provided analysis for estimating the word slope.

References