International Journal of Engineering & Technology, 7 (4.36) (2018) 1423-1427

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Skew detection based on vertical projection in latin character recognition of text document image

Ronny Susanto, Farica P. Putri*, Y. Widya Wiratama

Universitas Multimedia Nusantara, Tangerang, Indonesia
**Corresponding author E-mail: farica@umn.ac.id*

Abstract

The accuracy of Optical Character Recognition is deeply affected by the skew of the image. Skew detection & correction is one of the steps in OCR preprocessing to detect and correct the skew of document image. This research measures the effect of Combined Vertical Projection skew detection method to the accuracy of OCR. Accuracy of OCR is measured in Character Error Rate, Word Error Rate, and Word Error Rate (Order Independent). This research also measures the computational time needed in Combined Vertical Projection with different iteration. The experiment of Combined Vertical Projection is conducted by using iteration 0.5, 1, and 2 with rotation angle within -10 until 10 degrees. The experiment results show that the use of Combined Vertical Projection could lower the Character Error Rate, Word Error Rate, and Word Error Rate (Order Independent) up to 35.53, 34.51, and 32.74 percent, respectively. Using higher iteration value could lower the computational time but also decrease the accuracy of OCR.

Keywords: Optical Character Recognition, Preprocessing, Skew Detection, Projection Profile, Vertical Projection.

1. Introduction

Optical Character Recognition (OCR) has been an active research area since 1950 [1]. OCR is the capability of a machine to read and transform text document images into other medias and formats. OCR is one of the most successful technology application and has been used in various areas such as script recognition, banking, passport authentication, and language identification [2].

OCR is divided into online recognition and offline recognition based on the input device. Online recognition uses devices like digitizer tablets for data acquisition and the recognition is done while writing on the device. In contrast to online recognition, offline recognition uses devices like scanners and cameras to acquire data [3]. Offline recognition needs further processing of its data, also called preprocessing, which includes noise removal [2] and skew detection & correction [4] prior to character recognition process.

One of the most important steps in preprocessing of OCR is skew detection & correction. The goal of skew detection & correction is to deskew the image which is unavoidable when the image is scanned. The skew detection & correction is categorized into four category which are Hough transform, projection profile, nearest neighbor clustering, and interline cross correlation [4].

The most direct approach is projection profile which is first introduced by Postl [5] and is based on horizontal projection profile. The projection profile is calculated in some different angles. Using this approach Chauduri and Pal [6] introduced a skew detection technique to use in Hindi scripts. The technique is based on the fact that 32 of 50 Bangla characters and 42 of 49 Devanagari characters have a horizontal line on top [6].

In 2011, a novel skew detection & correction technique is introduced by Papandreou and Gatos. The technique is based on vertical projection profile. The use of vertical projection instead of horizontal projection is based on the fact that 33 of 52 Latin alphabets have at least 1 vertical line. Papandreou and Gatos further enhanced the accuracy of Vertical Projection Profile technique by combining it with the Minimum Bounding Box approach. The projection profile technique is calculated from different angles in a particular range. The projection profile methods are usually calculated within ±10 to 15 degrees [4].

Previous researches are conducted to detect the skew of non-Latin characters using profile projection [5], [6], [7]. However, in this paper, we study the implementation of skew detection based on profile projection in Latin characters. We extend our research to measure the impact of Combined Vertical Projection to the accuracy of OCR. The OCR accuracy is measured in Character Error Rate (CER), Word Error Rate (WER), and WER with Order Independent [8]. The computational time needed is also measured in 3 different iteration of Combined Vertical Projection technique.

2. Optical character recognition

The process of OCR consists of several steps as shown in Figure 1. Each step passes its result to the next step and there is no feedback loop to allow a process in the earlier step to use results in the later steps [1].

1. Data acquisition: the process of scanning and acquiring image in BMP, JPG, etc.
2. Pre-processing: pre-processing in OCR is needed to modify the image with the goal to fix the deficiency of the device used to acquire the image. Pre-processing is also needed to remove parts of image which are not intended to be recognized, such as watermark, uneven background, or other noises which are not needed in feature extraction. The expected result of pre-processing is a binary image that contains only text [7]. The steps of pre-processing are:
   a. Binarization: a process to transform grayscale image into binary image. The value of pixel in binary image is de-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

noted as $b_i \in \{0,1\}$. 0 represents black pixel and 1 represents white pixel.

b. Skew detection & correction: a few angles of skew is unavoidable while scanning a document. Skew in the image will decrease the accuracy of line segmentation in OCR and deeply affect the accuracy of OCR [9].
c. Segmentation: a process to split characters of an image into sub images of individual characters. This process includes line segmentation, word segmentation, and character segmentation [1]. Line segmentation splits the lines of text, sub images acquired from line segmentation is split into sub images of individual word by word segmentation, and are split into sub images of individual characters by character segmentation [10].
d. Feature extraction: this step analyses a text segment and selects a group of features to be used to identify the text [10].
e. Classification: the result of feature extraction is matched to a set of prototype character representing of each possible class. The distance between each pattern and prototype is computed. The prototype with the best match is assigned to the pattern [1].

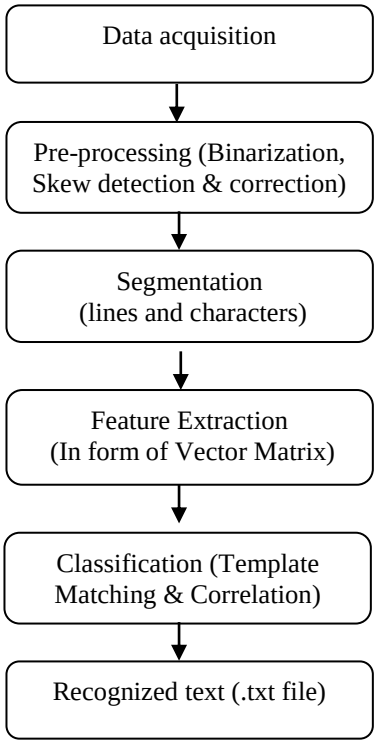


Fig. 1: Optical character recognition steps

3. Combined vertical projection

Projection profile analysis is a skew detection method of scanned document image. Projection is a process of that converts a binary image into one-dimensional array [11]. There are two types of projection profile: horizontal and vertical. Number of locations in the one-dimensional array equals to number of rows, for horizontal projection profile, or columns, for vertical projection profile, in the image. Each location in the projection profile stores a count of the number of black pixels in the corresponding row or column of the image, also called histogram [12]. This method is a straightforward solution to determining the skew angle of a document image uses a horizontal or vertical projection profile [12].



Fig. 2: Latin alphabets with vertical strokes circled

Combined vertical projection is a further enhancement of vertical projection method which was proposed by Papandreou and Gatos in 2011. It combines both minimum bounding box approach and the vertical projection profile technique. This method seeks higher concentration in some columns when the image is correctly aligned. This is due to the characteristic of most (33 out of 52) Latin alphabets having at least one vertical line as shown in Figure 2 [4].

Skew angle can be found by summing the black pixels of each column and compute the energy of an image rotated within a certain range. Higher concentration of black pixels in certain columns results in higher energy. Energy of an image rotated at θ angle can be computed using (1).

$$\sum_{i=1}^{m} C_i^2(\theta)$$

m is the width of image, and $C_i^2(\theta)$ is the square of black pixels count in i^{th} column.

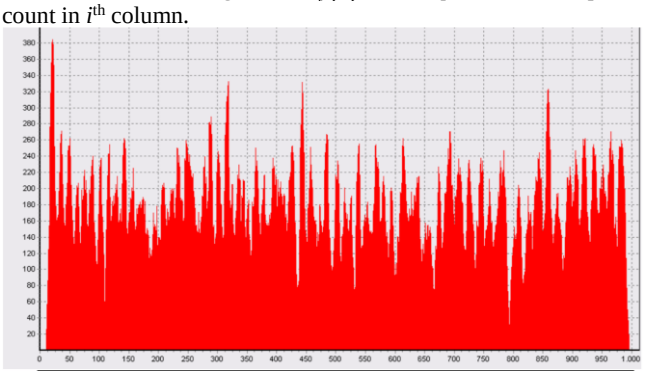


Fig. 3: Histogram of document image without skew

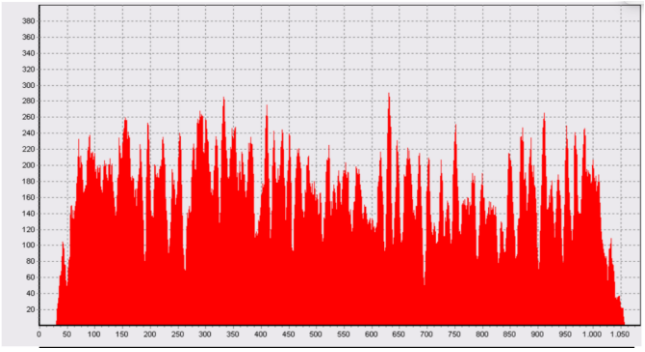


Fig. 4: Histogram of document image with skew

Figures 3 and 4 are the histogram of document image without skew and with skew, respectively. The peak in histogram shows the black pixels in each column on the document image. In Figure 3, when the document is properly aligned, the great peaks are observed, which shows a great concentration of black pixels. Figure 4 shows all the peaks will be average when the document is not properly aligned.

Minimum bounding box approach can be used to improve the accuracy of vertical projection skew detection. Minimum bounding box which includes every black pixels in the image is estimat-

ed and used to divide the energy calculated in (1) as the area of bounding box is minimum when the image is properly aligned as shown in Figure 5.

With the area of bounding box estimated. Energy of the image can be computed with (2).

$$T(\theta) = \frac{A(\theta)}{0.3\,(X_a - X_b)(Y_a - Y_d)} \quad (2)$$

X_a , X_b, Y_a, and Y_d are the extreme points of bounding box and 0.3 is a specified weight used to adjust the contribution of each technique. Skew angle of the image can be calculated with the maximum of $T(\theta)$ [4].

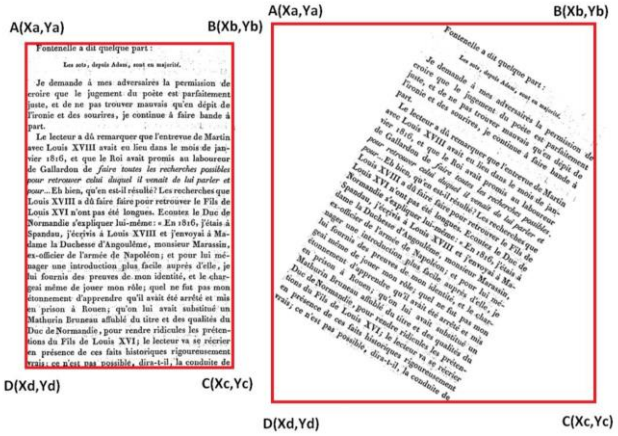


Fig. 5: Minimum bounding box approach

4. Experiment and result analysis

The experiment is conducted using rotation angle within -10 to 10 degrees. To recognize texts in the document image, Tesseract library is also used. Skew detection with combined vertical projection is done after image binarization which is done by Tesseract. Bounding box is found by scanning for first black pixel of each side. To measure the computational time needed on each data, DateTime function is used to capture interval between the start and the end of combined vertical projection technique.

In order to measure the impact of combined vertical projection, experiments are done on 40 document images acquired by scanners with manually skewed. The experiment includes measuring OCR accuracy on testing data without skew detection and measuring OCR accuracy using combined vertical projection with iteration value of 0.5, 1, and 2.

4.1. OCR without skew detection

As in Figure 6, OCR without skew detection on dataset produces an average error rates of 42%, 46.21%, and 42.69% in CER, WER, and WER Order Independent, respectively. Some tests indicate an error rate near 100%. Extreme skew of image causes texts in the image to be unrecognizable.

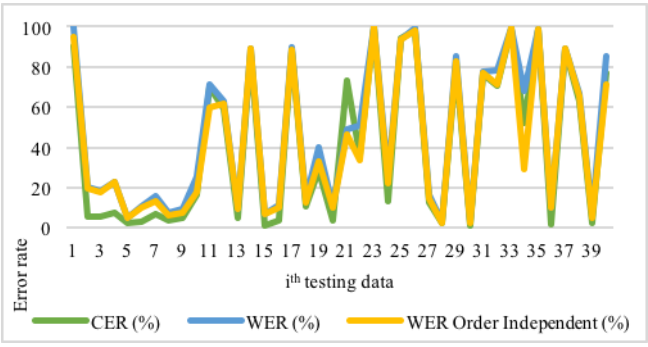


Fig. 6: Error rates of OCR without skew detection

4.2. OCR with combined vertical projection using iteration value of 0.5

Average error rates produced by OCR with Combined Vertical Projection (0.5 iteration) is significantly lower to error rates without a skew detection technique. Average error rates produced are 5.47%, 11.70%, and 9.95% in CER, WER, and WER Order Independent, respectively as shown in Figure 7. Figure 8 shows the required computational time measurement produced an average value of 326.19 seconds.

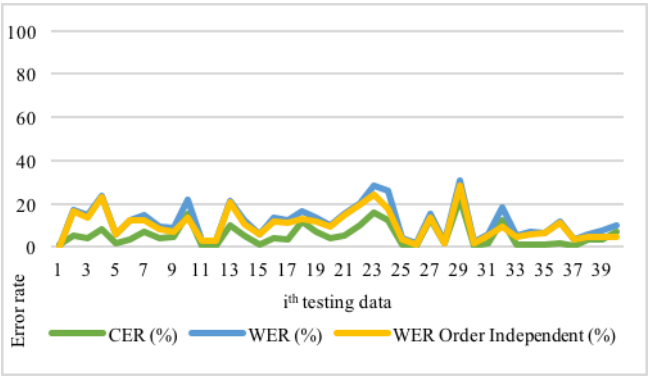


Fig. 7: Error rates of OCR with combined vertical projection using iteration of 0.5



Fig. 8: Computational time of OCR with combined vertical projection using iteration of 0.5

4.3. OCR combined vertical projection using iteration value of 1

The results of OCR with combined vertical projection using 1 as iteration value is shown in Figure 9. Error rates produced in this experiment does not show a significant change compared to using 0.5 as its iteration value. Average error rates produced are 5.86%, 11.90%, and 10.51% in CER, WER, and WER Order Independent. Figure 10 shows that average value of needed computational time is 171.49 seconds which is almost half of computational time needed in combined vertical projection using 0.5 iteration.

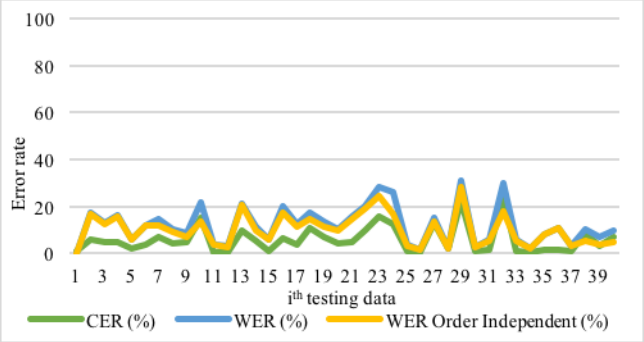


Fig. 9: Error rates of OCR with combined vertical projection using iteration of 1

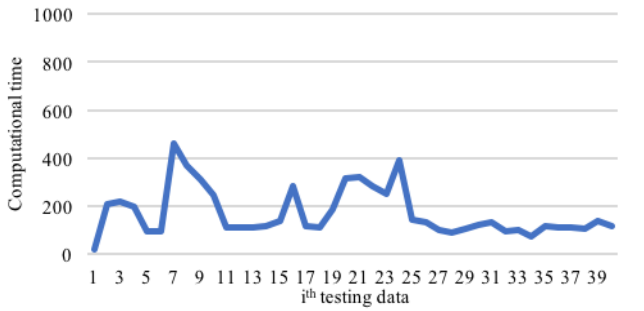


Fig. 10: Computational time of OCR with combined vertical projection using iteration of 1

4.4. OCR with combined vertical projection using iteration value of 2

OCR with Combined Vertical Projection using iteration value of 2 produced average error rates of 14.09%, 19.34%, and 17.37% in CER, WER, and WER Order Independent as shown in Figure 11. Test results on some data produced a near 100% of error rate. This proves iteration value of 2 is not reliable as it will occasionally fail to detect the image skew. Figure 12 shows the average computational time needed in iteration value of 2 is 88.08 seconds.

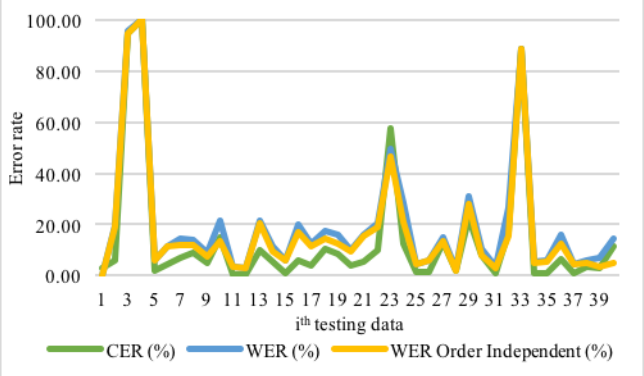


Fig. 11: Error rates of OCR with combined vertical projection using iteration of 2

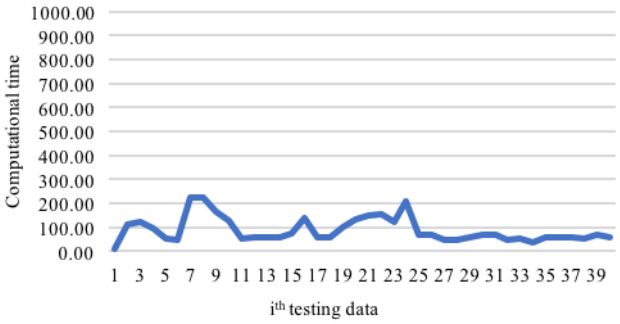


Fig. 12: Computational time of OCR with combined vertical projection using iteration of 2

Figure 13 is the average error rates of all the experiments describe before. It shows the combined vertical projection approach could lower the error rates of CER, WER, WER (Order Independent) up to 35.53, 34.51, and 32.74 percent. The lowest error rate is produced by using vertical projection with iteration of 0.5 because it calculates the energy every 0.5 degree of iteration. Thus, it can detect the skew up to 0.5 degree.

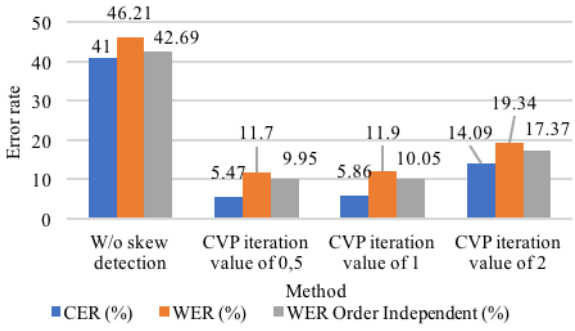


Fig. 13: Average error rates

The iteration value used in the combined vertical projection method is inversely proportional to the computational time. Figure 14 shows the use of iterations 2 times larger can also reduce the computational time about 2 times. It happens because the larger degree of iteration, resulting in more number of iterations.

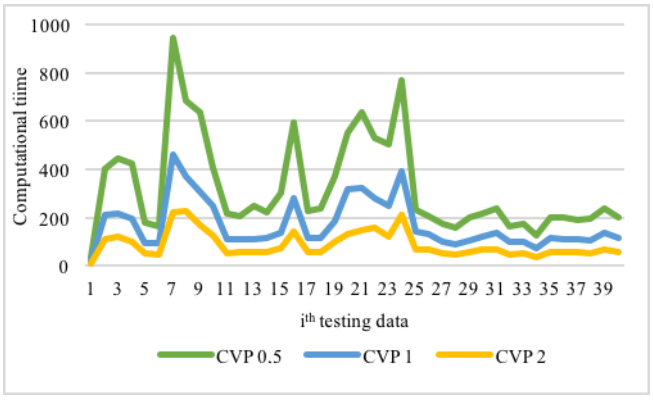


Fig. 14: Computational time of combined vertical projection with different iteration

5. Conclusion

Implementation of Combined Vertical Projection skew detection technique in OCR could improve OCR accuracy. Test results of OCR with Combined Vertical Projection produced lowest average error rates of 5.47%, 11.70%, and 9.95% using iteration value of 0.5, whereas error rates produced without skew detection technique are 41%, 46,21%, and 42.69% in CER, WER, and WER Order Independent. Average computational times needed by Combined Vertical Projection are 326.19 seconds, 171.49 seconds,

and 88.08 seconds using iteration of 0.5, 1, and 2 respectively. The use of higher degree of iteration value results in lower computation time. For further research, noise reduction or removal could be used to increase the accuracy of skew detection with combined vertical projection approach.

References

[1] Chandarana J & Kapadia MR, "Optical character recognition", International Journal of Emerging Technology and Advanced Engineering, Vol. 4, No. 5, (2014), pp. 219-223.

[2] Minoru M, Character Recognition, IntechOpen, (2010).

[3] Berchmans D & Kumar SS, "Optical character recognition: an overview and an insight", Proceedings of International Control, Instrumentation, Communication and Computational Technologies (ICCICCT), (2014), pp: 1361-1365.

[4] Papandreou A & Gatos B, "A novel Skew Detection technique based on Vertical Projections", Proceedings of International Document Analysis and Recognition (ICDAR), (2011), pp: 384-388.

[5] Postl W, "Detection of linear oblique structures and skew scan in digitized documents", Proceedings of International Conference on Pattern Recognition, (1986), pp: 687-689.

[6] Chauduri BB & Pal U, "An improved document skew angle estimation technique", Journal of Pattern Recognition Letters, Vol. 17, No. 8, (1996), pp. 899-904.

[7] Kant AJ & Vyavahare AJ, "Devanagari OCR using projection profile segmentation method", International Research Journal of Engineering and Technology, Vol. 3, No. 7, (2016), pp. 132-134.

[8] Carrasco RC, "An open-source OCR evaluation tool", Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, (2014), pp: 179-184.

[9] Smith R, et.al., "Tesseract Open Source OCR Engine", (2017), available online: https://github.com/tesseract-ocr/tesseract

[10] Vijayarani S & Sakila A, "Performance comparison of OCR Tools", International Journal of UbiComp (IJU), Vol. 6, No. 3, (2015), pp. 19-30.

[11] Al-Khatatneh A, Pitchay SA, & Al-qudah M, "A Review of Skew Detection Techniques for Document", Proceedings of International Conference on Modelling and Simulation (UKSim), (2015), pp: 316-321.

[12] Jain B & Borah M, "A survey paper on skew detection of offline handwritten character recognition system", International Journal of Computer Engineering and Applications, Vol. 6, No. 1, (2014).

[13] Poovizhi P, "A study on preprocessing techniques for the character recognition", International Journal of Open Information Technologies, Vol. 2, No. 12, (2014), pp. 21-24.