A Novel Skew Detection and Correction Approach for Scanned Documents

Conference Paper · April 2016

CITATIONS
14

READS
3,271

2 authors:

Andreas Dengel
Deutsches Forschungszentrum für Künstliche Intelligenz
837 PUBLICATIONS **9,383** CITATIONS
SEE PROFILE

Riaz Ahmad
Shaheed Benazir Bhutto University, Sheringal
25 PUBLICATIONS **427** CITATIONS
SEE PROFILE

A Novel Skew Detection and Correction Approach for Scanned Documents

Riaz Ahmad*, S. Faisal Rashid*, M. Zeshan Afzal*, Marcus Liwicki†, Andreas Dengel*, Thomas Breuel‡
*{rahmad, sheikh_faisal.rashid, afzal, Andreas.Dengel}@dfki.uni-kl.de, DFKI, TU-Kaiserslautern, Germany
† marcus.liwicki@unifr.ch, University in Fribourg, Switzerland
‡tmb@iupr.com, TU-Kaiserslautern, Germany

***Abstract*—Document analysis is relying on pre-processing stage. As much as we get optimal results in pre-processing, better will be the chances of analyzing the contents of a document. Similarly, for scanned text documents, a skew detection and then correction is very important, because, it significantly improves the accuracies in Optical Character Recognition (OCR) systems. This paper presents a novel approach for skew detection and correction. The approach is specifically tested on Latin and Arabic like scripts and outperform state of the art approaches.**

Keywords: Skew Detection, Scanned Documents, OCR

(a) Before skew correction

(b) After skew correction

Figure 1: A skewed document before and after skew correction.

1. Introduction

In any OCR system, correct extraction of text lines and their recognition mainly rely on skew detection and correction. A skew detection and correction are shown in Figure 1. The latest research regarding skew detection and correction can be found in; [1] [2] and [3].

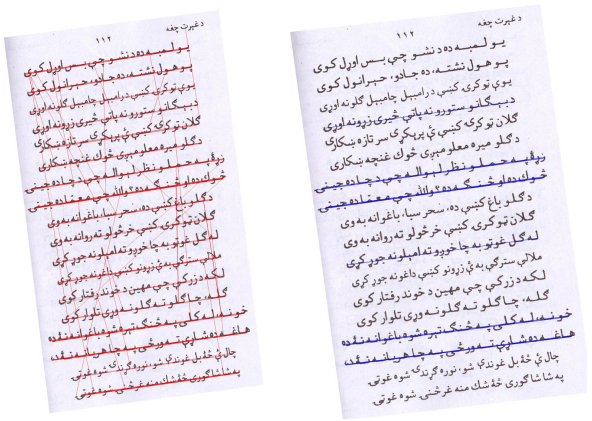
In general, skewed documents containing text are usually corrected after noise removal and proper binarization. This is followed by (1) detect lines in Region of Interest (RoI), (2) find the angles of these lines with respect to x-axis, (3) find the exact skew angle from angles of extracted lines and (4) to finally achieve the de-skewed image, rotate the image by exact skew angle. Technically stage (2) and stage (3) are important for skew detection and correction. Because, many lines can be detected in scanned document. However, the exact lines (say it true-lines) which measure the diverted/ skewed angle from x-axis are very important. Once we find these true-lines, it is very easy to find the exact skew angle.

The proposed approach is focused on stage (2) and stage (3). The novelty of our approach is the suppressing of detected false-lines (all other lines except true-lines) and then give weightage to the most probable true-lines. This is achieved by grouping detected lines in clusters. The lines, which are parallel to each other, will go to one cluster. Finally, the cluster having largest number of lines, represents a set of true-lines. From the set of true-lines, just one line having maximum length is considered for skew detection. The proposed approach is tested on real scanned images taken from Tobacco800[1] [4] and Pashto text. Recently Pashto language has been exploring towards the development of an OCR system [5] [6] [7] [8]. Thus, the proposed method is independent to different scripts like Latin and Arabic. We recommend our method for scanned documents. Because, it is efficient and reliable compare to other approaches, provided the scan documents with the skew range in between $\pm 30^{\circ}$ degrees.

2. Proposed Method

Similar to other approaches our proposed approach also works in fours stages (as mentioned in introduction). However, in stage (1) *"lines detection"*; Probabilistic Hough Transform (PHT) [9] is used, which is comparatively faster than simple Hough Transform (HT). In stage (2)*"True-lines detection"*; All the detected lines are grouped in relevant clusters w.r.t their parallelism with each other. Thus the cluster having largest number of lines, is considered as a representative of true-lines. In stage (3)*"exact skew angle is identified"*; The line having maximum length from the largest cluster is selected. The angle of the selected line w.r.t x-axis is a target skew angle. (4)*"de-skewed the image by rotation"*; the image is rotated by exact the skew angle. Step wise detail is mentioned in the following sub section.

1. The Legacy Tobacco Document Library (LTDL), University of California, San Francisco, 2007. [Online]. Available: http://legacy.library.ucsf.edu/



(a) true-lines + false-lines (b) Only true-lines

Figure 2: (a): Red lines represent all lines identified by PHT, (b) the only true-lines in blue color.

TABLE 1: Success rate and skew correction time of our method.

Datasets	De-Skewed	Success Rate%	Average Time (s)
Pashto Images	199/200	99.5	0.58
Tobacco800	48/50	96	0.98

2.1. Step-by-step Description of our Method

Step 0: Apply PHT to extract all possible lines, see Figure 2, (a).
Step 1: Lines are grouped into clusters w.r.t their same *gradient or slope* value. If a line has two points (x_1, y_1) and (x_2, y_2), then its gradient or slope *"m"* can be computed using equation 1, providing $(x_1 ! = x_2)$.
Step 2: The cluster having the largest number of lines is taken as a cluster of true-lines. In Figure 2, (b) the set of true-lines are shown.
Step 3: From the chosen cluster, the line having maximum length is selected and the angle of that line w.r.t x-axis is computed, which is the skew angle.
Step 4: Rotate the image by identified skew angle.

$$m = (y_2 - y_1)/(x_2 - x_1) \tag{1}$$

3. Results and Discussion

The proposed method is evaluated by using 200 scanned images of Pashto text and 50 images from Tobacco800 image database. All the images are scanned at 300 dpi. A PC is used with Intel Core2 Duo CPU P8600 @ 2.40GHz x 2, and with 2 GB of memory. The tests are performed in Python.

Results show that the overall success rate is 98.8%, and the average time taken by our method to de-skew an image is 0.78 second. In this context our proposed method gives better results both in terms of success rate and average time. The results are shown in Table 1. According to the work mentioned in [3], our results so far are the best results. However, due to dissimilar datasets we could not compare these results directly with the ones presented in [3]. The reason is the lack of information, particularly that what images are exactly used from certain dataset for skew detection and correction.

The images, which are chosen from Pashto text, differ in contents, which ultimately proves the versatility of our proposed approach and provide reliability to handle Latin as well as Arabic like scripts. Further, the specs of our PC is lower than the PC used in [3]. Therefore, we believe, that the average time for skew detection might be less than axis-Parallel Bounding Box method (APB) [3].

4. Conclusion

We have presented a novel approach for skew detection and correction in scanned documents. The proposed approach is based on the heuristic that text lines in a documents are always parallel to each other. Thus the cluster having maximum number of lines, can give us skew angle. Our method is time efficient, due to use of PHT. And avoid the use of some morphological operations like rotation before the detection of the correct skew angle. This saves much time in computing skew angel of the document. It is empirically shown through real scanned documents, that our proposed method is suitable for Latin as well as Arabic like scripts.

References

[1] Al-Shatnawi, Atallah M. "A skew detection and correction technique for Arabic script text-line based on subwords bounding". Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on. IEEE, 2014.

[2] Patel, Jinal, Anup Shah, and Hetal Patel. "Skew Angle Detection and Correction using Radon Transform." International Journal of Electronics, Electrical and Computational, System, ISSN 2348-117X (2015).

[3] Shafii, Mahnaz, and Maher Sid-Ahmed. "Skew detection and correction based on an axes-parallel bounding box." International Journal on Document Analysis and Recognition (IJDAR) 18.1 (2015): 59-71.

[4] Lewis, David, et al. "Building a test collection for complex document information processing." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

[5] Ahmad, R., Amin, S. H., Khan, M. A. (2010, October). Scale and rotation invariant recognition of cursive Pashto script using SIFT features. In Emerging Technologies (ICET), 2010 6th International Conference on (pp. 299-303). IEEE.

[6] Ahmad, Riaz, et al. "Scale and Rotation Invariant OCR for Pashto Cursive Script using MDLSTM Network." Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. IEEE, 2015.

[7] Ahmad, Riaz, et al. "Recognizable units in Pashto language for OCR." Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. IEEE, 2015.

[8] Ahmad, Riaz, et al. "Robust Optical Recognition of Cursive Pashto Script Using Scale, Rotation and Location Invariant Approach." PloS one 10.9 (2015): e0133648.

[9] Stephens, Richard S. "Probabilistic approach to the Hough transform." Image and vision computing 9.1 (1991): 66-71.