Skew Correction and Text Line Extraction of Arabic Historical Documents

Abdelhay Zoizou(✉), Arsalane Zarghili, and Ilham Chaker

Faculty of Sciences and Technologies, USMBA-Fez, Fes, Morocco
{abdelhay.zoizou,arsalane.zarghili,
Ilham.chaker}@usmba.ac.ma

Abstract. The field of optical character recognition for the Arabic text is not getting much attention by researchers comparing to Latin text. It is only in the last two decades that this field was being exploited, due to the complexity of Arabic writing and the fact that it demands a critical step which is segmentation; first from text to lines, then from lines to words and finally from words to characters. In case of historical documents, the segmentation is more complicated because of the absence of writing rules and the poor quality of documents. In this paper we present a projection-based technique for the segmentation of text into lines of ancient Arabic documents. To override the problem of overlapping and touching lines which is the most challenging problem facing the segmentation systems, firstly, pre-processing operations are applied for binarization and noise reduction. Secondly a skew correction technique is proposed beside a space following algorithm which is performed to separate lines from each other. The segmentation method is applied on four representations of the text image, including an original binary image and other three representations obtained by transforming the input image into: (1) smeared image with RLSA algorithm, (2) up-to-down transitions, (3) smoothed image by gaussian filter. The obtained results are promising and they are compared in term of accuracy and time cost. These methods are evaluated on a private set of 129 historical documents images provided by Al-Qaraouiyine Library.

Keywords: Arabic text line segmentation · Historical documents · Projection · Skew correction

1 Introduction

Since the beginning of research on optically captured text recognition of Arabic printed and handwritten text, the work on historical documents is still poor and not much researches could be found in literature, although there is a huge number of ancient documents in libraries and archives. These ancient documents contain important records either artistic, scientific, or referring to historical events. These files have not been digitally exploited yet, because the processing of ancient text is much different and harder than the processing of modern printed or handwritten text, especially for the Arabic text. The tasks of segmentation and recognition are more complicated due to the nature of Arabic writing and variety of shapes, where in most of Arabic ancient text there is a lot of additional signs and diacritics as shown in Fig. 1. The Arabic writing is

K. Smaïli (Ed.): ICALP 2019, CCIS 1108, pp. 181–193, 2019.
https://doi.org/10.1007/978-3-030-32959-4_13

done from right to left and each character can have different shapes depending on its position in the word Fig. 2.

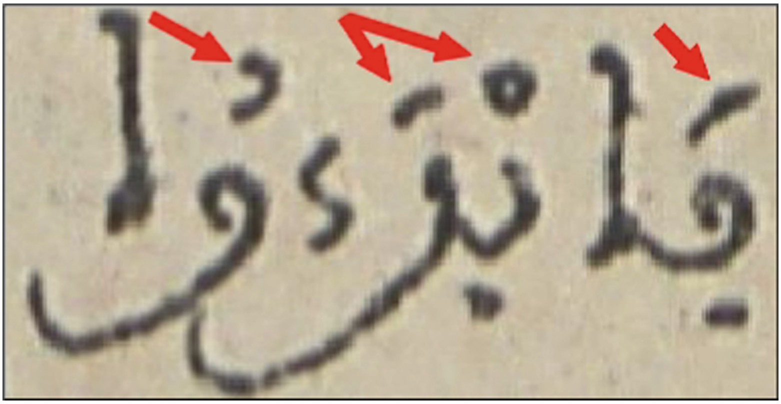


Fig. 1. Additional diacritic signs

Following the analytical approach, the recognition of an Arabic text must include a major step which is the segmentation of text into small classifiable units. To do this, the text must be decomposed first into lines and then into words and finally into characters [1]. One of the most challenging tasks for all text line extraction systems is the segmentation of overlapping and touching lines especially when dealing with skewed text. In this paper, we describe first, the pre-processing operations to be applied to the text image. Then, we propose a projection-based process to correct the skew of the text and to extract text lines by finding the path separating overlapping and touching lines.

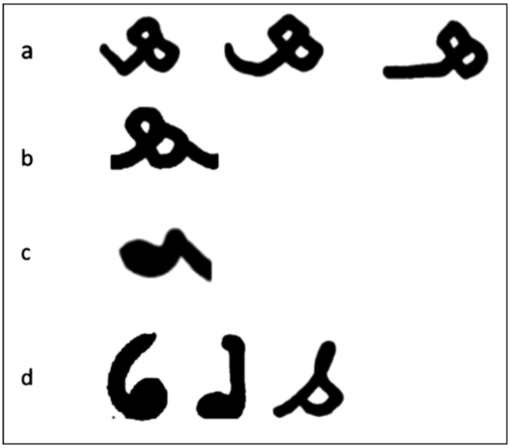


Fig. 2. Shapes of the same letter (هـ). a: in the beginning, b: in the middle, c: in the end, d: isolated

Historical documents have always presented a challenge to the segmentation systems, not only on the first level of text to line segmentation but also for line to words and word to characters. As handwritten, ancient text is often skewed and many additional signs could be found (Fig. 1), and it is written randomly with the pen thickness that changes frequently from one word to another.

OCR systems had always the challenge to find the right segmentation path between handwritten lines, either for Arabic or other languages. Multiple approaches have been adopted to deal with line segmentation problem; in [2], an overview of works on segmentation of Handwritten Document was presented. Authors of [3] proposed a language independent method for text line extraction. The Partial projection method was adopted in [4–6]. It is based on calculating the projection profile in multiple columns of the document. The clustering approach was adopted in many works on segmentation [7–9]. In [10], An algorithm based on adaptive local connectivity map was proposed for line extraction from historical documents. Authors in [11] developed a new cost function that considers the interactions between text lines and the curvilinearity of each text line. Recently, the problem of text line extraction starts having attention by deep learning researches, where in [12], deep convolutional network has been used for the segmentation of text lines of historical Latin documents.

The lines of text in Arabic historical documents are very close to each other and the distance between lines may also change many times in the same document. All these facts give the effect of overlapping and touching lines as indicated in Fig. 3. Lines overlapping effect occurs when ascenders and descenders of two successive lines share the same horizontal position. While touching line effect occurs when a descender of the first line meets an ascender of the second line. Another common challenge between Arabic Optical processing systems dealing with ancient documents is the absence of standard databases, although there is a huge number of historical documents in libraries.

This paper is structured as follows: some previous works are presented in Sects. 2, and 3 describes the contributions of this work; First the algorithm of skew correction is presented. Then, the four presentations of the text image are described and the process of line extraction is explained. Finally, the experimental setup and discussion of the results are presented in Sect. 4 and the conclusions and future work in Sect. 5.

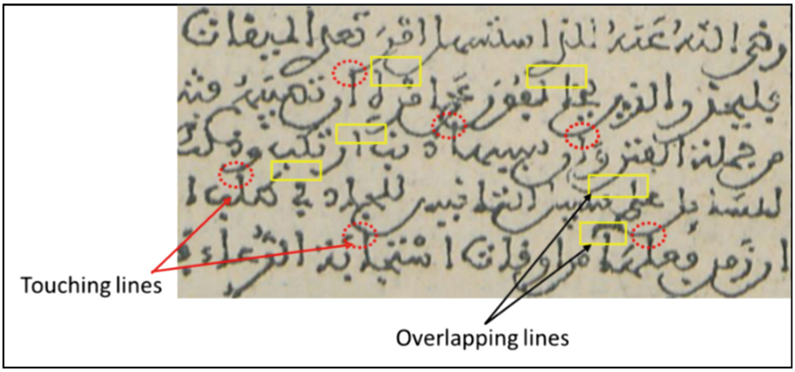


Fig. 3. Overlapping and touching lines

2 Previous Works

The task of text line extraction has been discussed since the very beginning of the OCR systems. The text line extraction is an important task in the process of segmentation which is the second step of a character-based recognition system. The text line extraction process depends on the nature of text, we found several methods applied for printed text as well as for the handwritten text. In 2001, authors in [4] proposed a method for text line segmentation, it is applied on modern handwritten text. The method is based on a combination of partial projection technique and partial contour. Authors claimed having a segmentation success rate of 82%, but the method fails in case of touching lines and the presence of diacritics signs.

Since then many approaches have been developed to deal with the special case of historical documents either for Latin [2, 3, 7, 8, 10], Chinese [9, 11], or Bangla [5]. Although, the work on the Arabic documents is still not getting much attention. In order to segment Arabic text lines, authors in [6] segment a page into columns, and they separate each column into blocks by using partial projection. Then, they use a KNN algorithm to classify each block into three classes depending on its dimensions: a class of big blocks representing overlapping components, a class for diacritic signs, and a class for words. The method has been tested for both printed and handwritten Latin text and only printed Arabic text, but no final success rate was given. In [13], a method for text line extraction from historical Arabic was developed. Its process goes first by detecting multi-skewed zones using an automatic paving and the Winger-Ville Distribution. Then the orientation of each zone and the baseline are analyzed to extract lines. Authors claim having a segmentation rate of 98.6%. This method is still weak in case of false inclinations and false maxima which is due to diacritics appearing in the beginning of the line.

Recently, machine learning techniques have been adopted to line extraction. In [12] authors used a method based on Convolutional Neural Network for Latin text line extraction from historical documents. Firstly, a patch size is identified on the basis of an overall text line distance. The CNN is fed for training with several patches. Then the CNN model is used for line segmentation. This method has been evaluated by using two sets of Latin databases and no results have been revealed. The use of CNNs is still unavailable in case of Arabic historical text because of the absence of a standard database of Arabic historical documents.

3 The Proposed Method

Text line segmentation is the operation of extraction of a text line separately with all its words and characters. In this work, the segmentation is done by using a projection profile-based method. The projection profile of a text image in a particular direction refers to the running count of the black pixels in that direction [14]. In our case, we use a horizontal projection.

The text in historical documents is often skewed because it is written by hand and the tools used did not help to write straightly on the line. For Arabic, writing was more difficult because lines were being written too close to each other with many ascenders, descenders and additional signs. This produce often the problem of more overlapping and touching between lines. The problem that was always the main challenge facing segmentation systems either for Latin or Arabic text. Even the most accurate systems haven't found a complete solution for the problem of overlapping and touching lines.

To address the line segmentation problem of historical Arabic documents, we propose a method based on projection first to correct skew angle of the whole page text at once, and then to separate lines. Before applying the segmentation and skew correction process, the text image is first being subject to some pre-processing operations including: Binarization and noise reduction. The method is tested on four different representations of the text image (Fig. 4). These representations were obtained by applying three transformations: smearing with RLSA algorithm, conversion into up-to-down transitions, and finally a gaussian smoothing.

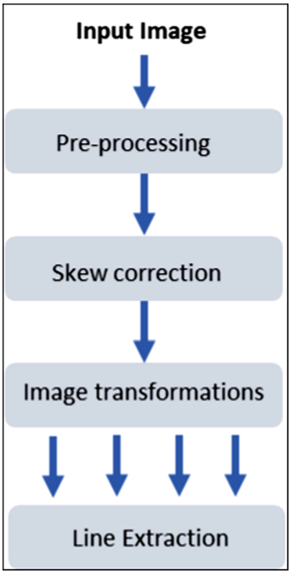


Fig. 4. The process of segmentation

3.1 Skew Correction

Skew correction problem has been discussed since decades. Many techniques have been used since then. But Hough Transform [15] and projection [14] are still the most used and efficient. Our skew correction process is based on horizontal projection which has been chosen to provide simplicity and efficiency and also for its rapidity. The algorithm is based on the calculating of the maximum peak of the projection profile of the text image in multiple rotations. But the weakness of this method is known face to overlapping and condensed lines; When applying the algorithm of skew correction on the original binary image, a false angle may be found, because often the orthogonal of a condensed text image has more pixels than lines, so that when rotating the image and calculating the projection profile, the maximum peak corresponds to this angle (Fig. 5).

Therefore, the original binary text image is first transformed into its up-to-down transitions (Fig. 6b). For each column of the binary image, only the black-to-white transitions are kept. The main goals of this transformation are first, to override the problem of condensed lines, and then to reduce the processed information to speed up the process because of the fact that the up-to-down transitions represent the baseline position and consequently the skew angle could be obtained by processing only the transitions.

The transition image is rotated in a range of angles [-a, a], with "a" is the fixed angle value. In each rotation, projection profile is calculated and its maximum value is saved as shown in Fig. 7. The principle is that the highest peak will be obtained by best rotation angle which gives straight lines since horizontal projection is used.

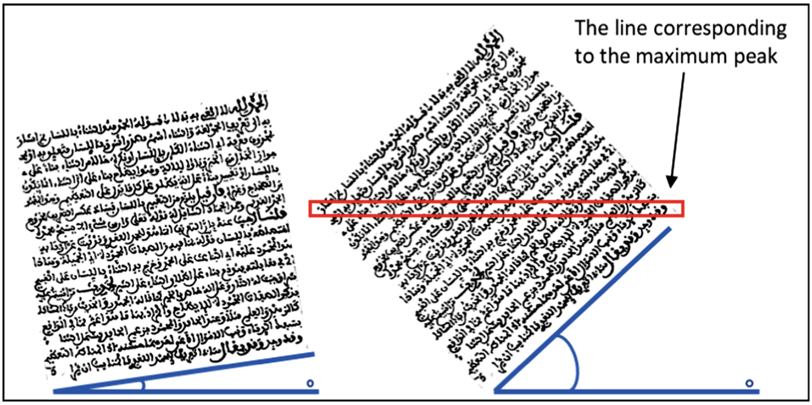


Fig. 5. False skew correction due to condensed and overlapped lines

3.2 Image Transformations

The process of segmentation is applied on four different representations of the text image. Three transformations have been applied on the text image (Fig. 8). The results of segmentation using each representation are compared to the original binary image (Fig. 8a). The first transformation is obtained by using Run length Smoothing Algorithm (RLSA) [16]. The algorithm is a smoothing technique that blacken the space between two black pixels in a single row or column. The white space is blacked if its size is less than a predefined threshold. It can be applied either on horizontal or vertical directions (Fig. 8b). The second transformation consists in converting the original binary image into its left-to-right and up-to-down black-white transitions. (Fig. 8c). The choice of keeping right transitions came from the fact that both Arabic writing and segmentation are done from right to left. While bottom transitions are chosen because they represent the baseline of the words and consequently the position of the line. The third and final transformation is done by using a binary gaussian filter. Similar to RLSA, the BGF transformation consists in applying a gaussian blur filter to add gray pixels between black pixels. Gaussian blur is a blurring technique based on a gaussian function which calculates the transformations to apply on a pixel. Then we apply a thresholding to convert the added gray pixel into black. (Fig. 8d).

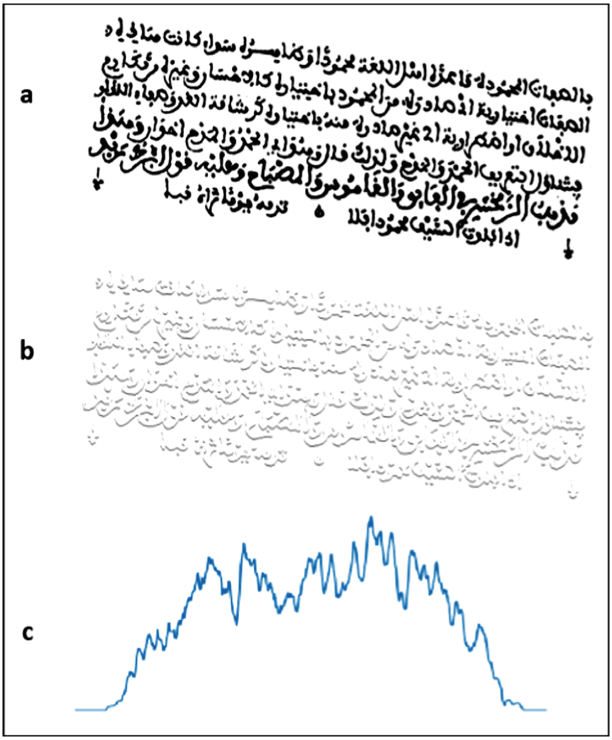


Fig. 6. Example of input of skew correction. a: original text image. b: its transitions transformation. c: its horizontal projection profile

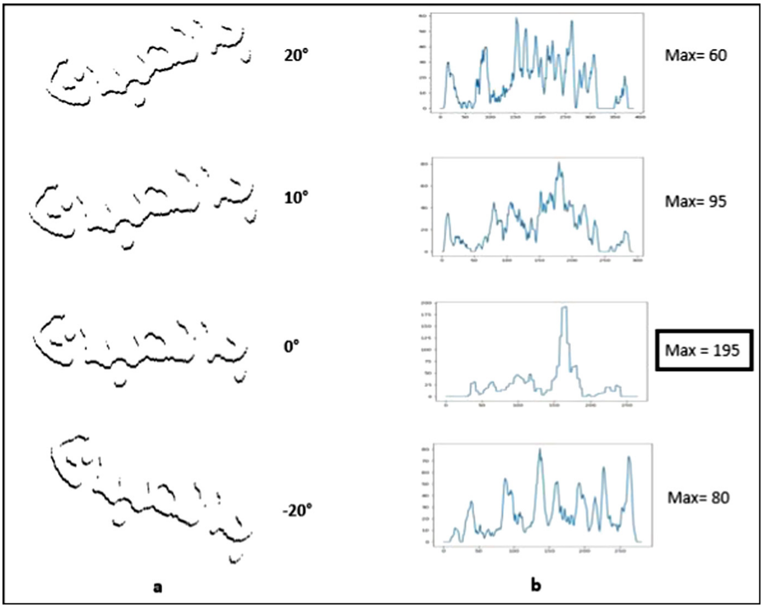


Fig. 7. Detection of the best angle for skew correction

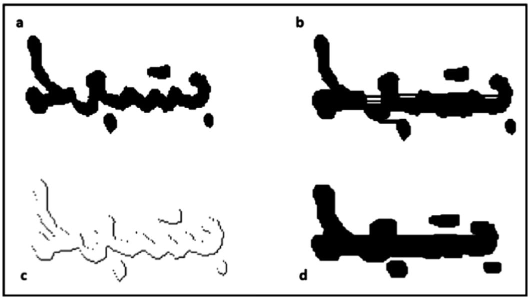


Fig. 8. Transformations applied to the text image. a: original image, b: transitions image, c: RLSA image, d: Binary gaussian transformation.

3.3 Text Line Segmentation

Text lines in historical documents are written too close to each other. Therefore, the obtained horizontal profile is noisy and lines' positions can't be found easily. To address this obstacle, we smooth the projection data vector using the SavGol filter [17]. Then local maxima representing the global line positions are detected and the segmentation area for each two consecutive lines is extracted, it is the area between the two

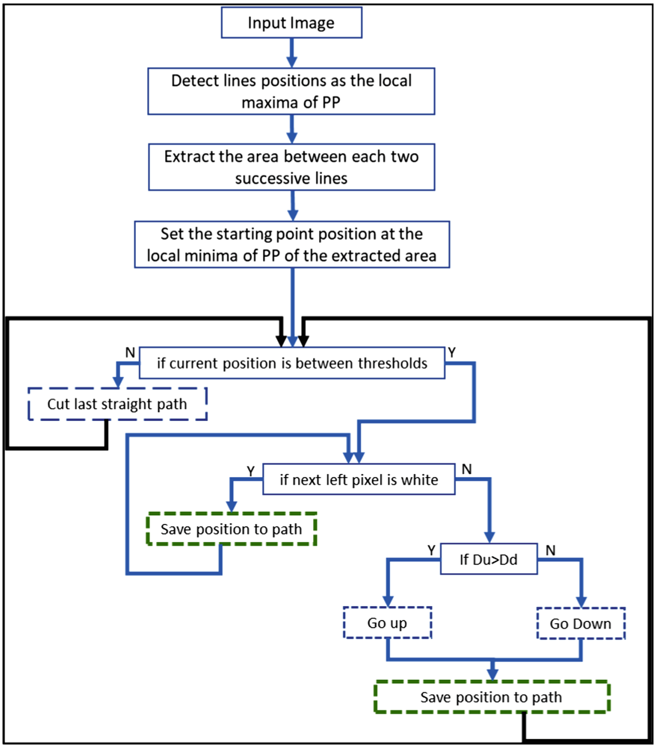

Input Image
Detect lines positions as the local maxima of PP
Extract the area between each two successive lines
Set the starting point position at the local minima of PP of the extracted area
N
if current position is between thresholds
Y
Cut last straight path
Y
if next left pixel is white
N
Save position to path
Y
If Du>Dd
N
Go up
Go Down
Save position to path

Fig. 9. Algorithm of lines segmentation

global baselines. Starting from the right index of local minima of the segmentation area, the algorithm of Fig. 9 is followed to find the segmentation path and separate lines from right to left. Let us mention that up and down thresholds are set to 20% of the segmentation area size, and Du (Dd) is the distance between the current pixel and its closest up (down) black pixel and PP is the projection profile. Figure 10 shows a segmentation area with the thresholds and the segmentation path obtained by using the Binary gaussian transformation.

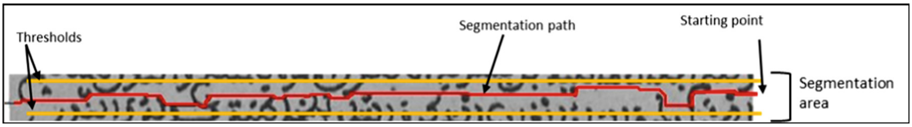


Fig. 10. Lines segmentation result

4 Experimental Results

Because of the absence of a public and standard database of Arabic historical documents, and in order to evaluate the present segmentation method, we have used a private set of Arabic historical documents. The documents were provided by Al-Qaraouiyine library[1], scanned with a 300 dpi optical scanner. The set contains a total of 129 pages with a total of 3068 text lines images. They were prepared manually to remove borders and to remove big black spots and the other effects that were caused by users during manual use. The tests were performed using Python-OpenCV on windows-10 of a PC with a Quadcore 2.8 Gz Processor and an 8 Gb of random-access memory.

4.1 Rules

The segmentation method described in this paper consists in finding a path through the white space between lines. In literature, we noticed the absence of standard scheme to decide if a line is well segmented or not. That's why we define the following strict rule to consider a successful segmentation: A line segmentation is checked manually and considered failed if the segmentation path crosses (cuts) one or more characters of the line in condition that a crossed character loses its global shape. Figure 11 shows some successful and failed segmentations.

[1] Al-Qaraouiyine Library: Founded in 860 in Fez, Morocco. Al-Qaraouiyine is believed to be the oldest working library in the world. It is part of Al-Qaraouiyine University which, according to the UN, is the oldest operating educational institute in the world.

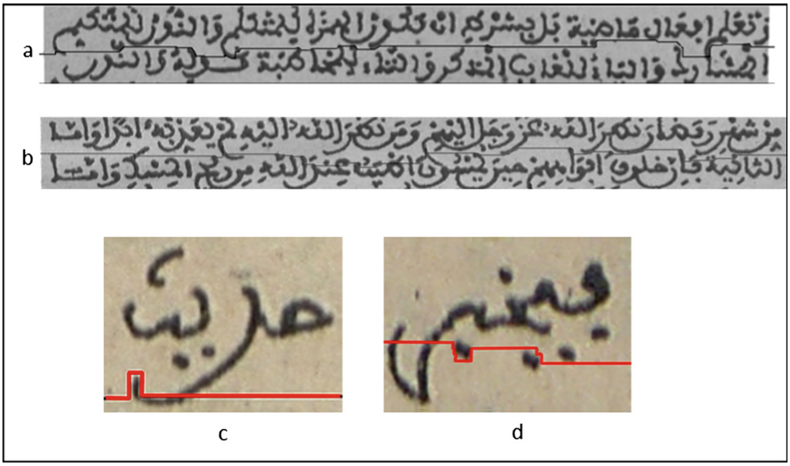


Fig. 11. a, c: Successful segmentations. b, d: failed segmentations

4.2 Skew Correction

In the skew correction step, the angle range is chosen experimentally as [−40°:40°] to cover all rotation variations that may exist in historical documents. The step angle of rotation is determined after several tests summarized in Fig. 12. We have chosen several divisions of the angle range, and calculated the error rate “r” using formula (1) and the time cost. We noticed that time increases whenever the number of steps increases since there are always new iterations to process, while the error rate of skew correction is stabilized at r = 1.8% at the number of steps of 70. Therefore, the range angle is divided by 70 which means the step angle is set to 80°/70 = 1.14°.

$$r = \frac{|ca - ea|}{ea} * 100 \tag{1}$$

Where ‘ca’ is the calculated angle, and ‘ea’ is the exact angle calculated manually.

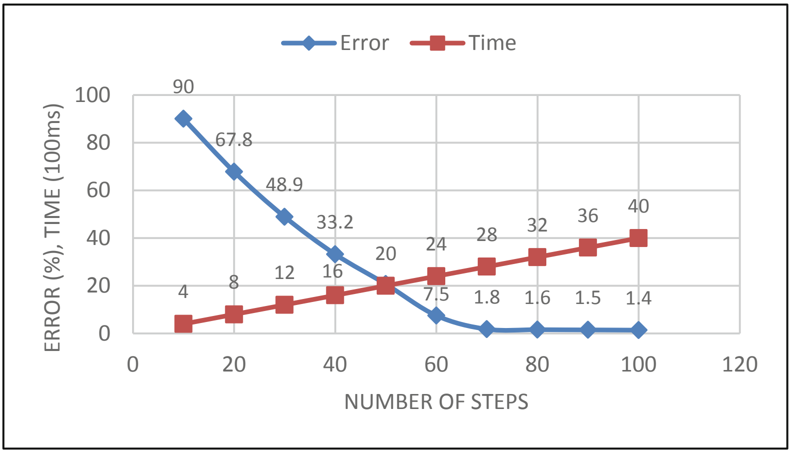


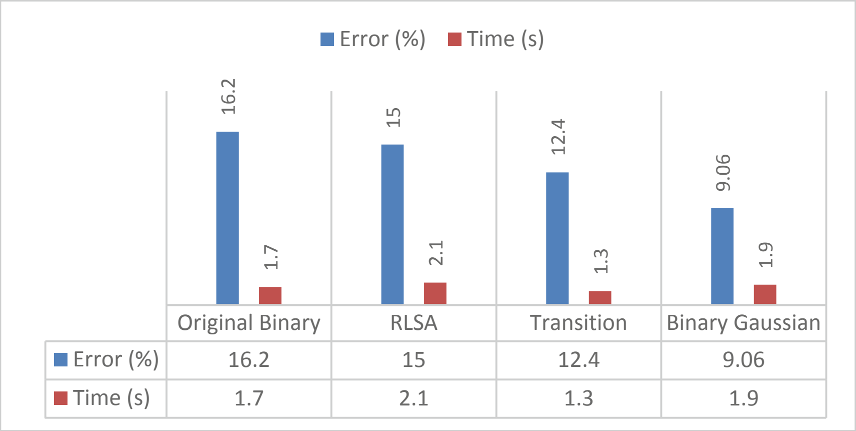
Fig. 12. Summary of skew correction tests

4.3 Text Line Segmentation

We present in Fig. 13 a comparison between the 4 previous tests in term of error rate and time consuming. The error rate is calculated using the formula (2). For the first representation in which the image is only binarized and got noise reduced, a global binarization algorithm is applied and a 2 × 2 kernel is used for a morphological 3-iterations opening to reduce noise. In the horizonal-RLSA representation, the threshold is experimentally fixed after many experiments in 15 pixels, the same as the kernel used for the gaussian filter.

$$e = \frac{|m - n|}{n} * 100 \tag{2}$$

Where 'm' is the number of lines well segmented, and 'n' is the total of lines in database.



	Original Binary	RLSA	Transition	Binary Gaussian
🟦 Error (%)	16.2	15	12.4	9.06
🟥 Time (s)	1.7	2.1	1.3	1.9

Fig. 13. Segmentation rates and time cost

In term of speed, processing the transitions image gives the best results since the transitions image have less data compared to the others. But the binary gaussian transformation gives the best results in term of accuracy nevertheless it takes more time than the original binary image and the transitions images. This is due to the fact that the binary gaussian transformation reduces noise, smooths the borders and converts the words into a smeared form, which is more compacted than the result of RLSA transformation. This last, smears the text image without reducing noise which can cause the loss of the segmentation path. While the transitions transformation creates more space and options for the segmentation path. Table 1 summarizes the obtained results.

Table 1. Results of segmentation using the four image representations

Representation	Segmentation rate
Original binary	83.8%
RLSA	85%
Transitions	87.6
Binary gaussian	**90.94%**

5 Conclusion

In this paper, we have presented a projection-based method for text line segmentation of historical documents preceded by skew correction. Since skew and noise are major weaknesses of projection-based approaches of text line segmentation, we first perform a preprocessing step including binarization and noise removing. Then, a skew correction method is proposed to provide straight lines to the segmentation process. It is based on the horizontal projection, which is applied on the transformation of the text image into transitions. The segmentation algorithm is applied on multiple versions of the text image to see the effect of each transformation. We have found that adding a gaussian blur followed by thresholding to a binary text image may be the best option and can increase the segmentation rate up to 90.94%. As perspective of this work, we aim to study the robustness of our method on bigger databases and other Latin and Chinese scripts. Also, we aim to create a standard database for Arabic historical text, to allow the application of new machine learning techniques in this field.

References

1. Zoizou, A., Zarghili, A., Chaker, I.: A new hybrid method for Arabic multi-font text segmentation, and a reference corpus construction. J. King. Saud. Univ. – Comput. Inf. Sci. (2018). https://doi.org/10.1016/j.jksuci.2018.07.003
2. Katsouros, V., Papavassiliou, V.: Segmentation of handwritten document images into text lines. Image Segmentation (2012). https://doi.org/10.5772/15923
3. Saabni, R., Asi, A., El-Sana, J.: Text line extraction for historical document images. Pattern Recognit. Lett. **35**, 23–33 (2014). https://doi.org/10.1016/j.patrec.2013.07.007
4. Zahour, A., Taconet, B., Mercy, P., Ramdane, S.: Arabic hand-written text-line extraction. In: Proceedings International Conference Document Analysis Recognition, ICDAR 2001-January, pp. 281–285 (2001). https://doi.org/10.1109/ICDAR.2001.953799
5. Pal, U., Datta, S.: Segmentation of Bangla unconstrained handwritten text. In: Proceedings International Conference Document Analysis Recognition, ICDAR 2003-January, pp. 1128–1132 (2003). https://doi.org/10.1109/ICDAR.2003.1227832
6. Boussellaa, W., Zahour, A., Elabed, H., et al.: Unsupervised block covering analysis for text-line segmentation of Arabic ancient handwritten document images. In: Proceedings - International Conference Pattern Recognition, pp. 1929–1932 (2010). https://doi.org/10.1109/ICPR.2010.475

7. Garz, A., Fischer, A., Bunke, H., Ingold, R.: A binarization-free clustering approach to segment curved text lines in historical manuscripts. In: Proceedings International Conference Document Analysis Recognition, ICDAR 1290–1294 (2013). https://doi.org/10.1109/ICDAR.2013.261
8. Garz, A., Fischer, A., Sablatnig, R., Bunke, H.: Binarization-free text line segmentation for historical documents based on interest point clustering. In: Proceedings- 10th IAPR International Work Document Analysis System DAS 2012, pp. 95–99 (2012). https://doi.org/10.1109/DAS.2012.23
9. Yin, F., Liu, C.L.: Handwritten text line extraction based on minimum spanning tree clustering. In: Proceedings 2007 International Conference Wavelet Analysis Pattern Recognition, ICWAPR 2007 3, pp. 1123–1128 (2008). https://doi.org/10.1109/ICWAPR.2007.4421601
10. Shi, Z., Setlur, S., Govindaraju, V.: Text extraction from gray scale historical document images using adaptive local connectivity map. In: Proceedings International Conference Document Analysis Recognition, ICDAR 2005, pp. 794–798 (2005). https://doi.org/10.1109/ICDAR.2005.229
11. Koo, H.L., Cho, N.I.: Text-line extraction in handwritten Chinese documents based on an energy minimization framework. IEEE Trans. Image Process. **21**, 1169–1175 (2012). https://doi.org/10.1109/TIP.2011.2166972
12. Capobianco, S., Marinai, S.: Text line extraction in handwritten historical documents. In: Grana, C., Baraldi, L. (eds.) IRCDL 2017. CCIS, vol. 733, pp. 68–79. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68130-6_6
13. Ouwayed, N., Belaïd, A., Auger, F.: General text line extraction approach based on locally orientation estimation. Doc. Recognit. Retr. XVII **7534**, 75340B (2009). https://doi.org/10.1117/12.839518
14. Casey, R.G., Lecolinet, E.: Survey of methods and STR in character segmentation. IEEE Anal. **18**, 690–706 (1996). https://doi.org/10.1109/34.506792
15. Likforman-Sulem, L., Hanimyan, A., Faure, C.: A hough based algorithm for extracting text lines in handwritten documents. In: Proceedings International Conference Document Analysis Recognition, ICDAR 2, pp. 774–777 (1995). https://doi.org/10.1109/ICDAR.1995.602017
16. Wong, K.Y., Casey, R.G., Wahl, F.M.: Document analysis system. IBM J. Res. Dev. **26**, 647–656 (2010). https://doi.org/10.1147/rd.266.0647
17. Savitzky, A., Golay, M.J.E.: Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. **2**, 1627–1639 (1964). https://doi.org/10.1021/ac60214a047