Text Segmentation of Historical Arabic Handwritten Manuscripts Using Projection Profile

Arwa Alghamdi, Dareen Alluhaybi, Doaa Almehmadi,
Khadijah Alameer, Sundos Bin Siddeq, Tahani Alsubait
College of Computer and Information Systems
Umm Al-Qura University
Makkah, Saudi Arabia
arwalghamdi@gmail.com, dareen.f.18@gmail.com, doaaahmaad7@gmail.com,
Dojalhashmi98@gmail.com, sundos.bs@gmail.com, tmsubait@uqu.edu.sa

2021 National Computing Colleges Conference (NCCC) | 978-1-7281-6719-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/NCCC49330.2021.9428836

***Abstract*—The research in the field of recognizing Arabic handwritten texts is flourishing with the passage of time. Despite that, progress in this field is considered insignificant and has many challenges, one of which is that the available Arabic datasets are limited. Also, due to the complexity of Arabic language, it is difficult to perform necessary preprocessing steps, such as lines and characters segmentation, especially characters segmentation since Arabic letters are usually overlapped into each other. In this study, we perform lines and characters segmentation based on Projection Profile methods (PP) which will be applied on a historical Arabic manuscript that was collected from Umm Al-Qura University's King Abdullah Bin Abdulaziz Library, Manuscripts department. The manuscript is a Moroccan Quranic manuscript that was written before more than 250 years ago. The results on lines segmentation are considered promising. On the other hand, poor results were achieved on characters segmentation due to the problem of complexity that was mentioned earlier.**

***Index Terms*—Arabic handwritten text, characters segmentation, image processing, lines segmentation, manuscripts, projection profile.**

I. Introduction

Throughout history, manuscripts were the official way of writing down knowledge and science. There are millions of manuscripts all around the world written in different languages. One of the most popular and important languages is the Arabic language. Despite that, the progress in Arabic handwritten text recognition research field, especially of historical documents, is still insignificant compared to other languages. There are many causes of this, one of which is the complexity of Arabic language which makes the problem more complicated and difficult to solve. Furthermore, since the manuscripts are ancient and written using old style of writing, this makes the problem even more complicated. An important initial step in text recognition is text segmentation, which segments the text into its basic components, i.e., lines, words, and characters [1]. In this study, we try to contribute to this important field by presenting an approach of segmenting historical handwritten Arabic manuscripts into lines and characters using Projection Profile method.

Segmentation is considered one of the most important steps for recognizing handwritten texts. If the segmentation is inaccurate, it will produce poor results in the recognition step. Handwritten text is a real challenge in text segmentation, since the segmentation is applied under the assumption that the text is written in a straight and consistent manner, which is mostly not true in handwritten text, because lines of text might be overlapped with the neighboring lines [1].

Simply put, segmentation is partitioning the image to detect different levels in images which are [1]:

1) Lines level.
2) Words level.
3) Characters level.

Each level will be discussed in more detail in the following sections.

A. Lines Segmentation

In lines segmentation, the image is segmented into lines. Each line consists of a group of words. It is achieved by scanning the image horizontally, starting from left top, until right bottom. This scanning is done in pixels level [1]. One of the most used methods in lines segmentation is baseline extraction. Baseline extraction tries to extract the baseline of each line in the text to perform the segmentation. It can be achieved using different methods such as Projection Profile (PP), nearest neighboring clustering, and Hough transform [1]. In this research, we will be focusing on using Projection Profile. Projection Profile method includes both Horizontal PP and Vertical PP [2]–[5].

B. Words Segmentation

Words segmentation is performed by finding the spacing between words. This is done by scanning each line vertically, from left top until right bottom. This is done in pixels level. Word segmentation can be achieved using several methods such as the Vertical PP [1], [5].

C. Characters Segmentation

The main goal of characters segmentation is to segment a word image into characters. It is usually a challenging task since it assumes there is a space between each character, which might also occur within a character. This problem leads to

inaccurate segmentation results since a single character might be split into two parts or more by mistake [1].

The rest of this paper will be organized as follows: related work were we discuss some of the similar studies, empirical work where we discuss our experiments details, results and discussion, and finally conclusion and future work.

II. Related Work

Segmentation methods in the related research studies are diverse. Moreover, the results differ depending on several factors such as the type of dataset, the overlap between characters and lines, and the noise which may affect the segmentation results.

One of the most popular methods used in segmenting texts into lines, words, or characters is the PP method. There are many studies that used this method to perform text segmentation. Moreover, there are some studies that used some clustering algorithms with connected component concept. All these studies will be discussed in the following sections.

In [2]–[10], the authors used the PP method to perform lines segmentation, most of them made some prepossessing steps such as binarization, noise removal and other steps in order to improve the work of the algorithm and to produce better results.

For example, in [6], the dataset used is called HADARA80P. The colored images were converted into grayscale images. After that, to facilitate lines detection, 2D images were converted into 1D signals. Then, PP algorithm was used.

Another study [7] used many Arabic manuscripts as a dataset. Firstly, the images were converted into grayscale before extracting the lines. In addition, in [10], some filters were applied on images to remove the noise, such as a convolution with a Gaussian filter.

The dataset in [2] is a dataset of Arabic manuscripts written by many scientists of Islam. Each image was converted into grayscale then into binary format. For another study [8], the Quran text was used, there was some noise in the images therefor it was removed using some smoothing processes.

Regarding actual segmentation, various methods were used in different studies. For example, in [2], [3], [5], [8] lines segmentation is done by using Horizontal PP.

In addition to the previous studies, three datasets were used in [9] which are: H-DIBCO 2016,H-DIBCO 2017, and OMAR-Database. The lines segmentation is done by using Vertical PP. Before that, document's images were smoothed to achieve a better segmentation.

Another study [11] used a new method of text segmentation. It works by finding the position of the baselines. After that, each line is segmented into separable characters depending on the baseline. The baseline is a horizontal line that contains the most numbers of pixels compared with all horizontal lines on the script.

In [12], [13], the dataset used for segmentation is a middle-age Persian documents. A set of 200 pages have been scanned from different volumes of "The Pahlavi Manuscript Collection". In both of them, the connected components approach was used.

In [13], connected component approach and morphological analysis have been used. Based on this idea, document lines are decomposed into many connected components. These components may be characters or words. Firstly, preprocessing and image restoration algorithms were applied such as thresholding and noise removal. Then, morphological connected component detection is used for detecting the line components using a certain equation. Finally, there are conditional steps and equations which were used to extract all text components. The results of text line extraction and removing other lines overlaps show an accuracy of 97.35%. And 99.5% accuracy for segmenting the extracted text lines to their components (words and characters).

On the other hand, in [12] a clustering algorithm was used with connected components concept. At first, the image is converted into binary, then it will be divided into connected components. There are points located in the middle of the connected component which will be taken as an input for the clustering algorithm. After that, these points are clustered using the k-means algorithm and the final lines will be produces. The k-means algorithm was chosen since it has proved its success on page to lines segmentation. Thus, clustering results will be the best choices for segmentation the page's lines. Finally, the accuracy of lines segmentation by using clustering algorithm was 98.12%..

There are other methods which were used in lines segmentation, such as RU-net in [14]. The final accuracy that was obtained in lines segmentation is 96.71%.

Moreover, another method was used in [15]. Firstly, the text is segmented into vertical slices, where the size of each block in X-axis is measured, when two blocks have a low value, all blocks within them will be considered as one block of text.

Word segmentation is carried out in different ways. In [5], [6], [8], the segmentation is achieved by using the Vertical PP after applying some processing steps.

In [5], to improve the segmentation, the input image was deskewed. This is done by calculating the skew angle using the spectrum of the input image. The spectrum will be segmented in four quadrants. After that, the angle for each one of them will be calculated. Then, the image will be rotated using the average of these angles' values. Finally, the segmentation will be applied on the text.

Furthermore, in [6] dilation operation was applied based on morphological filtering of the binary image along each segmented line. In [8] the noise was prior to the segmentation process as well as applying binarization, thinning, smoothing, normalization, and baseline detection processes on images.

In [3], words are segmented manually by the expert from the line and stored in the training set.

In another study [14], the CNN with BLSTM was used for words segmentation, and CTC was used to find the alignment between the text-line transcription and the text-line image. An accuracy of 80.1% was obtained for words segmentation.

In [7], related components are extracted and labelled using the binary images. Then, projection operation is applied at each line to produce the words. The main problem with this technique is its sensitivity to overlapping words and noise.

Characters segmentation in [2] is performed by using Vertical PP. Calculations of densities is performed for each column. This method will segment the characters if the Vertical PP density of each character in the line is less than the Horizontal PP.

III. Empirical work

In this section, we will be discussing our experiments on performing lines and characters segmentation using PP method on a historical Arabic manuscript written in Moroccan font.

In this research, we have used Python language for implementing the algorithms. This includes using different libraries. The main ones are OpenCV and scikit-image for applying the preprocessing on images, and Seaborn and Matplotlib for plotting purposes.

In order to reach excellent segmentation results, it is needed to apply some preprocessing steps on the manuscript. These steps will be discussed in the following subsections.

A. Images Segmentation

The original manuscript was in PDF format. Therefor, it was converted into individual PNG images. Since all images contain two pages except one image which contains only one page, each image was cropped into two images. After cropping, the total number of images was 61.

B. Noise Removal

This step focuses on removing any noise of the images which contain the text to be segmented.

Since the manuscript is ancient and written more than 250 years ago, pages contain a considerable amount of noise. Furthermore, an important characteristic of Arabic language is diacritics which can be added to letters in different locations, e.g., at the top or the bottom of the letter. These diacritics make the text segmentation process more complicated. As a solution to this, colored images of the manuscripts were split into three main channels: R, G, and B. After that, the R channel was chosen as it was the clearest one among them.

Another step of noise removal that was taken is applying Gaussian filter which is in charge of smoothing the image.

C. Binarization and Inverting

This step aims to convert the colored images into binary images where the text (foreground) is in white and the background is in black. This is required in the PP method.

Firstly, the images are converted into grayscale. Secondly, images are converted into binary by applying a specific thresholding technique. For this reason, different methods were applied on a sample of the manuscript in order to achieve the best results.

There are two main types of thresholding methods: global and local [9]. Global methods choose one threshold value for the whole image. On the other hand, local methods choose a different threshold value for each pixel by calculating the features of its neighbors [9]. Firstly, two local methods were applied which are adaptive Gaussian thresholding in Fig. 1a and adaptive mean thresholding in Fig. 1b. It is clear that both results are considered poor. So, other global methods were applied such as mean in Fig. 2a, Otsu in Fig. 2b, and minimum in Fig. 2c. It is clear that minimum method gives out the best result so it was the chosen method for binarization.

(a) Adaptive Gaussian (b) Adaptive mean

Fig. 1: Local thresholding methods results

After applying the binarization, the result was the text (foreground) in black color and the background in white color. In order to make the text become more clear, the invert step was performed where the text (foreground) has become in white color and the background in black color.

(a) Mean (b) Otsu (c) Minimum

Fig. 2: Global thresholding methods results

D. Margins Cropping

Each page in the original manuscript contains margins in three directions depending on the side of the written text. Therefor, it was needed to crop these margins so that the images will contain only text without empty spaces.

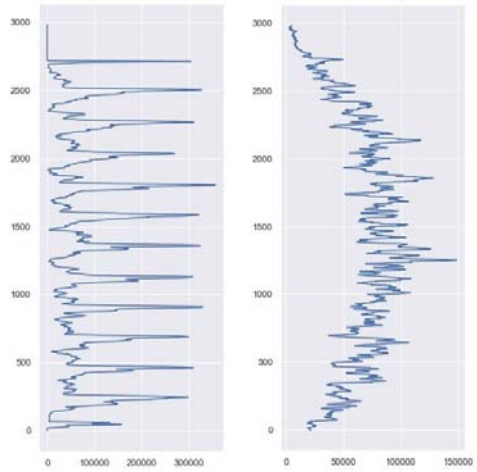
This process is composed of many steps. Firstly, inverted images are used as input, then the adaptive (local) threshold is calculated for each block, where the block size is set to 35 and the max value of threshold is set to 255. After that, erosion method is applied to the image using a structure element of size 40×40. Then, a mask is extracted from the erosion image by applying a global thresholding with a value of 120. The next step is that the edges of the images are detected and the max

and min values of both x and y axes are calculated. Finally, these values will be used to crop image's margins.

E. Skew Detection and Correction

Skew happens when the image is not set correctly on the scanner or the camera, which results in poor accuracy in the segmentation phase.

Some parts of the manuscript contain skewed images which will therefore affect the segmentation phase. Therefor, a skew detection and correction method was applied where the binary image is used as input. Then, different values of angles are tested in order to find the best angle value. This is done by calculating the Horizontal PP. The difference between each value and the value next to it powered by two is calculated. If the image is skewed, the sum of the differences between these two values will be small. In contrast, if the image's skew is set correctly, the difference between them will be large. To clarify this further, Fig. 3 shows two Horizontal PP, where Fig. 3a is the correctly skewed page, and Fig. 3b is the same page but after rotating it by 30°. Finally, the angle that leads to the maximum score will be chosen. After that, the image will be rotated using this angle value.



(a) Deskewed page (b) Skewed page

Fig. 3: The effect of skew on Horizontal PP

F. Segmentation

The segmentation step is dividing text images into base components such as lines, words, and characters [1]. In this paper, the work focuses on lines segmentation and characters segmentation.

1) Lines Segmentation: After the images' margins have been cropped, images are segmented line by line using the Horizontal PP method, which converts 2D into 1D image by calculating the densities of each row. Then, each line is cropped separately at the lowest density point [2]. Since the manuscript is written in a complex way, the spaces between lines is not totally empty. Therefor, its density from Horizontal PP is not equal to 0. In order to overcome this problem, an opening operation is performed on each image before segmentation. Opening is considered a morphology operation. Morphology operations are divided into two fundamental categories [13], [16]:

a) Erosion: The erosion operation uses a structuring element for reducing or shrinking the shapes contained in the input image:

$$A \otimes B = \{z \mid (B)_z \subset A\} \tag{1}$$

where A is the image, B is the structure element, and z is the points in B.

b) Dilation: The dilation operation uses a structuring element characteristics for expanding the image:

$$A \oplus B = \{z \mid (B)_z \cap A \neq \Phi\} \tag{2}$$

where A is the image, B is the structure element, and z is the points in B.

Opening is an erosion followed by a dilation using the same structuring element. By doing this, small elements will be shrink during erosion process. After that, the remaining elements will be grown back again during dilation process [17].

Even after doing opening operation, some empty lines were segmented. Thus, a condition was applied so that any line which its width is smaller than a specific threshold will be excluded.

2) Characters Segmentation: Another type of segmentation that was performed is characters segmentation. In this step, the lines that are obtained from the previous step are divided into separated characters.

This step was applied by using Vertical PP method. This method works the same way as Horizontal PP in the previous step, except that the direction of calculating the density. in Vertical PP it is calculated on columns instead of rows [2]. Opening operation was also performed on each line before segmentation. More details about opening operation is mentioned in III-F1.

Fig. 4 summarize the methodology used in the empirical work section.

IV. Results and Discussion

The total number of lines segmented during lines segmentation is 635. The original manuscript contains 657 lines. There was 10 empty segmented lines out of 635, since they contain parts of the characters from the line above and below them. Fig. 5 shows an example of this case. In addition, 15 lines were overlapped with other lines. Table I shows the results' summary of lines segmentation. Since the results of lines segmentation will affect characters segmentation phase, empty lines were excluded and overlapped lines were segmented manually. In total, 644 lines were obtained.

There are some difficulties in lines segmentation, especially when there are some characters under the lines.

Unlike lines segmentation, the results of characters segmentation is not very satisfying, as some characters were not cropped correctly for many reasons. Since Arabic characters

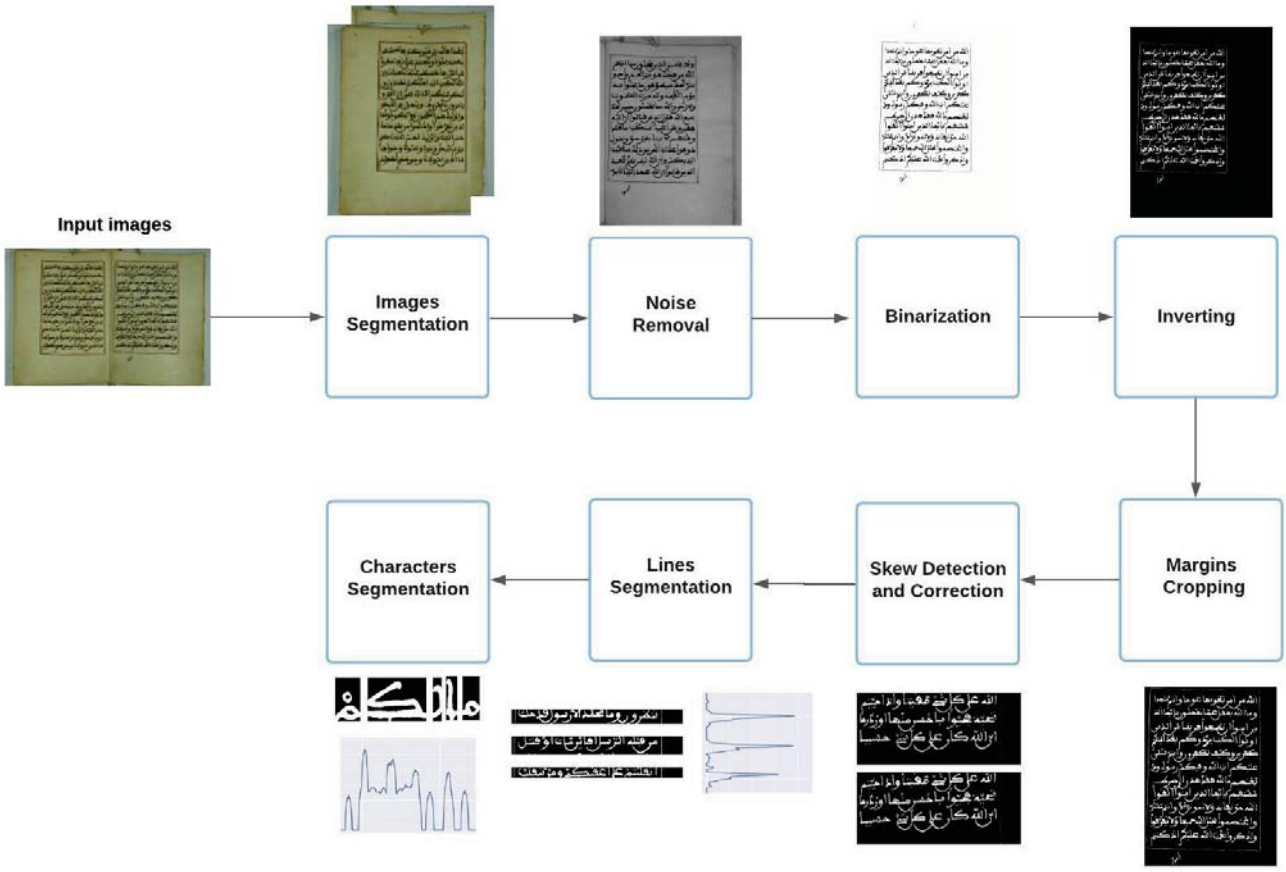


Fig. 4: Empirical work methodology

Fig. 5: Empty line

TABLE I: Lines Segmentation Results Summary

Case	Number of Lines
Total number of segmented lines	635
Correctly segmented lines	610
Empty lines	10
Two overlapped lines	11
Three overlapped lines	4

in one word may have several forms. For example, the same character might be either connected or separated from its adjacent characters. Fig. 6 shows an example of two forms for the same character. This makes the process of characters segmentation more difficult, as Vertical PP will not always have a value of 0 between characters.

The total number of characters after segmentation is 11,099. Where the original manuscript contains 16,199 characters. However, there was a lot of segmented characters that are either empty or overlapped with other characters.

(a) Separated character (b) Connected character

Fig. 6: Different character's forms

V. Conclusion and Future Work

In this research, we presented an approach for lines and characters segmentation based on PP method. This method was applied on a historical Arabic manuscript collected from the Manuscripts department at Umm Al-Qura University's Library. At the end of this work, we notice that good results were obtained in lines segmentation. In contrast, the results were not satisfactory in characters segmentation,. There are different plans to improve the results of the segmentation either by changing some steps of the preprocessing methods, or by adopting new techniques, such as using deep learning model to apply the segmentation instead of the PP method.

Acknowledgment

We would like to thank the Manuscripts Department in King Abdullah Bin Abdulaziz University Library for providing us the valuable manuscript.

REFERENCES

[1] N. Dave, "Segmentation methods for hand written character recognition," *International journal of signal processing, image processing and pattern recognition*, vol. 8, no. 4, pp. 155–164, 2015.

[2] B. Alrehali, N. Alsaedi, H. Alahmadi, and N. Abid, "Historical arabic manuscripts text recognition using convolutional neural network," in *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*. IEEE, 2020, pp. 37–42.

[3] L. Dounas, M. I. Azzouzi, and N. Rais, "New algorithm for the transcription of arabic manuscripts," in *2012 Colloquium in Information Science and Technology*. IEEE, 2012, pp. 86–90.

[4] A. Maqqor, A. Halli, K. Satori, and H. Tairi, "Using hmm toolkit (htk) for recognition of arabic manuscripts characters," in *2014 International Conference on Multimedia Computing and Systems (ICMCS)*. IEEE, 2014, pp. 475–479.

[5] A. A. Shinde and D. Chougule, "Text pre-processing and text segmentation for ocr," *International Journal of Computer Science Engineering and Technology*, vol. 2, no. 1, pp. 810–812, 2012.

[6] M. Elmansouri, N. E. Makhfi, and B. Aghoutane, "Toward classification of arabic manuscripts words based on the deep convolutional neural networks," in *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*. IEEE, 2020, pp. 1–5.

[7] N. E. makhfi, "Handwritten arabic word spotting using speeded up robust features algorithm," in *2019 5th International Conference on Optimization and Applications (ICOA)*. IEEE, 2019, pp. 1–6.

[8] A. Iqbal and A. Zafar, "Offline handwritten quranic text recognition: A research perspective," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*. IEEE, 2019, pp. 125–128.

[9] M. M. Dyla and F. Morain-Nicolier, "Text line segmentation and binarization of handwritten historical documents using the fast and adaptive bidimensional empirical mode decomposition," *Optik*, vol. 188, pp. 52–63, 2019.

[10] J. Ha, R. M. Haralick, and I. T. Phillips, "Document page decomposition by the bounding-box project," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 2. IEEE, 1995, pp. 1119–1122.

[11] Z. Osman, L. Hamandi, R. Zantout, and F. N. Sibai, "Automatic processing of arabic text," in *2009 International Conference on Innovations in Information Technology (IIT)*. IEEE, 2009, pp. 140–144.

[12] S. Alirezaee, H. Aghaeinia, K. Faez, and R. Rashidzadeh, "An efficient preprocessing block for the middle-age persian manuscripts," in *Canadian Conference on Electrical and Computer Engineering, 2005.* IEEE, 2005, pp. 2170–2173.

[13] S. Alirezaee, H. Aghaeinia, M. Ahmadi, and K. Faez, "An efficient restoration algorithm for the historic middle-age persian (pahlavi) manuscripts," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 3. IEEE, 2005, pp. 2114–2120.

[14] C. Neche, A. Belaid, and A. Kacem-Echi, "Arabic handwritten documents segmentation into text-lines and words using deep learning," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 6. IEEE, 2019, pp. 19–24.

[15] W. Boussellaa, A. Zahour, B. Taconet, A. Alimi, and A. Benabdelhafid, "Praad: Preprocessing and analysis tool for arabic ancient documents," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 1058–1062.

[16] C. R. Giardina and E. R. Dougherty, *Morphological methods in image and signal processing*. Prentice-Hall, Inc., 1988.

[17] K. A. M. Said, A. B. Jambek, and N. Sulaiman, "A study of image processing using morphological opening and closing processes," *International Journal of Control Theory and Applications*, vol. 9, no. 31, pp. 15–21, 2016.