ISSN 2464-4617 (print)
ISSN 2464-4625 (DVD)
Computer Science Research Notes
CSRN 2901
WSCG Proceedings Part II

Title Segmentation in Arabic Document Pages

Hassina Bouressace
University of Szeged,
13 Dugonics square,
Szeged, H-6720 Hungary
bouressacehassina@hotmail.fr

ABSTRACT

Recent studies on text line segmentation have not focused on title segmentation in complex structure documents, which may represent the upper rows in each article of a document page. Many methods cannot correctly distinguish between the titles and the text, especially when it contains more than one title. In this paper, we discuss this problem and then present a straightforward and robust title segmentation approach. The proposed method was tested on PATD (Printed Arabic Text Database) images and we achieved good results.

Keywords

Arabic language, Projection Profile, Text-line segmentation, Connected Components

1 INTRODUCTION

The goal of Document Analysis and Recognition (DAR) [Sim08] is the automatic detection and extraction the information existing on a page, where the output of DAR systems is usually presented in a structured symbolic that can then be processed by a computer. The principle of DAR is closely related to official documents such as (newspapers, business letters, books, and journals), where the information of these documents is presented in digital form such as a PDF, HTML or via a digital camera containing textual information.

The type of document structure may be a simple structure or complex one, where the identification of the structure is based on the amount of information contained in the document and the way this information is presented. A document with a complex structure is usually composed of textual heterogeneous blocks, which may contain mathematical expressions, tables, graphs and, pictures [Azo95]. They may be characterized by variability of positioning, shape, and appearance of areas, in which the different text blocks are not perfectly aligned, with a complex layout that can have multiple columns with different sizes, an irregular body, and spacing.

The two families of composite documents are complex with structurally stable documents (form, commercial letter, etc.) and complex with variable structure documents in which text is found between blocks and in others (newspapers, documents, magazines, flyers, etc.). They have rich typography and are not composed solely of text but a combination, in a variable arrangement, of texts, graphics, and images. Figure 1 shows two examples of images of documents with a complex structure. With the availability of the high-resolution scanning devices along with robust computers, the Optical Character Recognition (OCR) systems can handle numerous recognition tasks of text-images. Basic existing research used text-lines as input entities for an OCR system [Gra09], hence text-line extraction plays a key role in the OCR process and its facilitation. This has led to a lot of research on improving text-line segmentation over the years [Ray15, Ryu14].

(a) (b)

Figure 1: Examples of documents with a complex structure (composite): (a)A complex structure in stable document form, (b)A complex structure in variable document form (newspaper page).[Mon11]

The application of text-line segmentation is not always easy to do, due to the existence of skew, script variations, noise, text-lines with different sizes and different fonts. One especially problematic issue, which is the aim of the study, is the line segmentation of a large scale heading, in such a way that we can present it as titles and their subtitle detection in Arabic document pages. The existing approaches for text-line extraction cannot correctly distinguish the titles from the text, especially when it contains more than one

title. The real text in documents often contains titles and subtitles, and such text lines cannot be precisely identified with state-of-the-art methods.

The state-of-the-art approach described in [Mun17] could not extract Arabic text documents with large-scale headings and titles; moreover, it is not efficient in the case of a document with a complex structure. This inspired us to develop a new method that can extract not one title, but every title and subtitle on a document page. In this paper, we present a new text-line detection method for complex-structured documents where the detected text is treated as a title or subtitle and each page contains many titles corresponding to the article numbers. The paper is organized as follows: In Section 2, the related work is described. In Section 3, we describe each step of the algorithm in detail. Experiments and results are presented in Section 4 and finally, in Section 5 we outline our conclusions and plans for the future.

2 RELATED WORK

A wide variety of title detection methods for documents can be classified and incorporated in many techniques: Active Contour Model (Snake), Horizontal Projection Profile (HPP), Vertical Projection Profile (VPP), Connected Components (CCs), the Bounding box-based method, smearing method, the Hough Transform (HT), or use of HMMs. Here, the main studies of text-line detection methods are outlined.

Bukhari et al. [Buk08] presented a robust text-line segmentation approach against skew, curl and noise, which is based an active contour model (Snake) with the novel idea of several baby snakes and their convergence in a vertical direction using the ridges which are found by applying multi-oriented anisotropic Gaussian filter banks, hence it is computationally expensive. In [Zek11, Che01], they applied horizontal projection profile (HPP) and vertical projection profile (VPP) techniques for the text-line segmentation approach by finding the inter-line gap and taking into account the separation between two consecutive lines. In [Bou18], they applied a horizontal projection, commencing with a calculation of the histogram of each block to extract the local minima by using a threshold value, and conflict resolution for assigning the existing black pixels in the separator zones to the nearest line of text by a proximity analysis. However, this method is limited to specific structures. In [Sou10], they used the Bounding box-based method where a histogram is created, then they specified the lines that have the minimum number of pixels. Afterward, the boundaries of each line were detected by determining the centroid by measuring the regional properties. In [Hus15, Alj12, Bro13], they applied a smearing method; namely smearing the consecutive black pixels in the horizontal projection, then the pixels between them were marked in black if the distance between any two was less than a threshold value. However, it fails when there is no space between two consecutive lines or overlapping lines.

3 DESCRIPTION OF THE METHOD

Titles are the key elements of documents because there are no page documents without titles and subtitles. The size of these titles is not always larger than other text on the page, especially when the title belongs to a small article. Nevertheless, subtitles are usually found above or below the title where the space and the size between the subtitle and the article text are identical. These generic characteristics present challenges in the Arabic language in terms of text-line extraction from a document page. Figure 2 illustrates the problem where the spaces between the peaks did not give us useful information for title extraction.


في تقرير رفعه إتحاد الفلاحين إلى رئيس الجمهورية

الولاة، رؤساء الدوائر والأميار وراء نهب الأراضي الفلاحية!

■ ولاة العاصمة، عنابة وهران وبومرداس على رأس قائمة المتهمين

رفعت الأمانة العام للإتحاد الوطني للفلاحين، تقريرا أسود إلى الرئيس، عبد العزيز بوتفليقة، من خلال رئيس مجلس الأمة، تشكو فيه التجاوزات والتلاعب الحاصل من طرف السلطات المحلية عبر كافة ولايات الوطن وتناشده من أجل التدخل العاجل لوضع حد لمثل هذه التجاوزات.

حبيبة محمودي

3.1 Pre-processing

We used global thresholding to produce a clear image that simplifies the processing of the later stages. In this stage, the Otsus binarization method [Zek11] is used to transform the image into two possible pixel values (0 and 1) to reduce the noise and overcome the illumination issue that arises during the scanning process.

3.2 Removing Figures and black blocks

We know of course that document pages may contain one or more figures. These figures consist of the largest proportion of pixels in some cases that give us imprecise information and this could lead to poor results in the subsequent steps. Moreover, the existence of black blocks could corrupt the essential parts that are needed later on. To overcome these problems and facilitate title and subtitle segmentation, we used formulas applied constraints on the size of the connected components, the ratio of height and width, and the density of black pixels in the connected component. In Figure 3, we give an example of figure deletion and black block reverse.

ISBN 978-80-86943-38-1

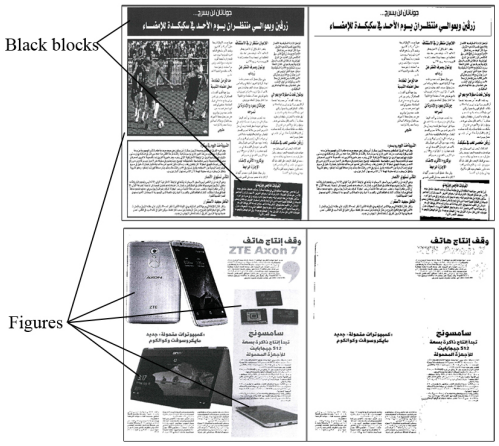

Figure 3: Examples of removing figures and black blocks.

3.3 Title Segmentation

The detection of the titles is done by taking into account the fact that not just the height of the titles is important but also the number of pixels in each component and its position. Here, our proposed method is based on RLSA and the Connected Components (CCs) technique. HRLSA, or the (Horizontal Run length smoothing algorithm)[Gor97] is then applied to the resulting image of the preceding step to eliminate spaces between words of the same line of text and Vertical RLSA smoothing is used to connect the diacritic marks to the corresponding words. Let L_0 be a horizontal segment of unit length. The Run-Length Smoothing closing algorithm fuses nearby pixels of the binary image X by γ L_0, where γ is a size parameter.

$$RLSA(X) = X \oplus \gamma L_0 \ominus \gamma L_0 \qquad (1)$$

The horizontal and vertical smoothing thresholds were determined empirically, namely (with threshold 1=1%) and (with threshold 2=0.85%) proportional to the size of the page, respectively. Actually, the characters of the titles are usually larger than those of the lines of simple text, in this case, the threshold of the horizontal RLSA smoothing was previously not enough to connect the words of a big title. To remedy this problem, we applied a second horizontal RLSA smoothing with a larger threshold (with threshold 3= 1.55 % proportional to the size of the page) only on the parts of the image containing probable major titles. These are composed of connected components whose height is greater than (1.5 x the most common text height in the document). We then applied another labeling of the related components on the RLSA smoothed image. As the words of a single line of text (simple or title) become connected, each line of text is treated as a separate component. A related component is treated as a title if its height is greater than (1.2 x the most common text height in the document), otherwise, it is treated as a simple line of text.

Figure 4: The title segmentation results of our proposed method on Arabic text documents.

3.4 Subtitle Extraction

However, these techniques only provided us with the main titles, no subtitles. Other criteria must be used to add the other titles. For this, we combined two criteria, namely the size of the related component of the previous step and its position relative to the main titles. Here, the other titles are extracted using the projection profiles (PP) method.
Let L_1 and L_2 denote the lines of text that are above and below a main title T respectively, V_1 denote the image width and V_2, V_3 denote the height. And let:

- $V_1 = 12.5\%, V_2 = 3.35\%, V_3 = 0.07\%$.
- (x_1, y_1): the coordinates of the top left-hand corner of L_1.
- (x_2, y_2): the coordinates of the bottom right-hand corner of L_1.
- (z_1, k_1): the coordinates of the top left-hand corner of T.
- (z_2, k_2): the coordinates of the bottom right-hand corner of T.
- (x_3, y_3): the coordinates of the top left-hand corner of L_2.
- (x_4, y_4): the coordinates of the bottom right-hand corner of L_2.

ISBN 978-80-86943-38-1

The lines of text L_1 and L_2 are treated as subtitles if they satisfy the following conditions:

- height of L_1 and $L_2 <$ (Threshold 4) 1.15 % proportional to the size of the page document;
- $(z_1 - x_2 < V_1) \wedge ((|\, y_1 - k_1 \,| < V_2) \vee (|\, y_2 - k_2 \,| < V_2))$
- $((k_1 - y_1 > V_3 \vee y_2 - k_2 > V_3) \wedge (y_1 - k_1 > V_3 \vee k_2 - y_2 > V_3))$
- $(x_3 - z_2 < V_1) \wedge ((|\, k_1 - y_3 \,| < V_2) \vee (|\, k_2 - y_4 \,| < V_2))$
- $((y_3 - k_1 > V_3 \vee z_2 - y_4 > V_3) \wedge (k_1 - y_3 > V_3 \vee y_4 - k_2 > V_3))$

Figure 5: The subtitle segmentation results for an Arabic document page.

We proceed in the same way with other subtitles if they exist by letting $(x1, y1)$ be the coordinates of the upper left-hand corner of T, (x_2, y_2) be the coordinates of the lower right-hand corner of T, (z_1, k_1) be the coordinates of the upper left-hand corner of subtitle L, (z_2, k_2) be the coordinates of the bottom right-hand corner of subtitle L in a recursive way until no line satisfies these conditions.

The conditions applied for subtitle and title detection are determined by the following geometrical features:

- Height: CC bounding box height,
- Width: CC bounding box width,
- Aspect Ratio: Width divided by height,
- Solidity: Area of the CC (in pixels) divided by the area of its convex hull,
- Area: Number of pixels in the CC,
- Position: CC coordinates.

Figure 6 shows the number of document pages in each threshold where the threshold is computed by getting image information under valid conditions. Every threshold in this chart is used for title or subtitle detection, which has four tests (with proportional values of 0.85, 1, 1.15 and 1.55, where these values are found by calculating the median of proportional values of the images (the ratio is extracted using information about the dimension and size of the document image, all the titles and subtitles being on each individual page). Here, the y parameter denotes the number of images that matched this proportional value. We took the best and the highest number column for each threshold. Our results were tested on over three hundred pages to check accuracy and performance.

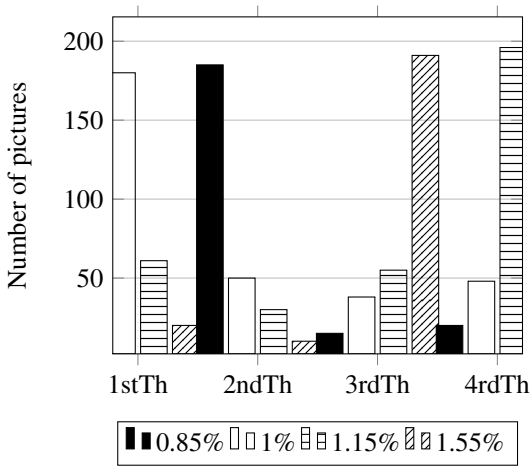

Figure 6: The number of experiments of our data thresholds.

4 EVALUATIONS AND RESULTS

Our method can be used in two modes; namely, the application returns just the cropped titles and subtitles, or it returns the whole page with colored titles and subtitles. Both demonstrate the segmentation phase in a clear way.

To evaluate our proposed system, we used the same criterion as in that described in [Ari07], which is the title-segmentation accuracy in percentage terms. The constraints are given below.

1) If a single connected component of a line is segmented to another line, this error if counted as two line errors.

2) If n subtitles and simple text are merged together, then it is counted as n line errors.

3) If n titles and subtitles are merged together, then it is counted as n line errors.

We used the following formula for computing all the errors:

$$Accuracy\% = 100 - (E / TotalTitleLines) * 100 \quad (2)$$

ISBN 978-80-86943-38-1

The algorithm was tested on three hundred scanned pages at 300 dpi got from the PATD (Printed Arabic Text Database) [Bou19], which has various styles including regular and bold, multiple font sizes and types(AL-Quds, AxTManal, Beirut, AL-Quds Bold, Kacstone, Alshrek Titles). Pages with different structures may contain one article, two articles, three articles or more. Every page has a unique format because the PATD database was collected from many documents. The algorithm gives excellent scores, which may be as high as 98.02% for titles, and 98.15% for subtitles. Table 1 below lists the results obtained during the testing process with various font types, styles, and sizes.

Table 1: Test results.

Font Type	Title extraction	Subtitle extraction
AL-Quds	98.18 %	97.96 %
AxTManal	98.15 %	98.23 %
Beirut	/	98.87 %
AL-Quds Bold	98.45 %	98.17 %
Kacstone	97.56 %	97.55 %
Alshrek Titles	97.78 %	/
Total	98.02 %	98.15 %

Due to the absence of previous articles with the same goal in Arabic documents, we evaluated the performance of our approach by comparing it with related articles that have similar goals such as line segmentation. In [Mun17], they took a binarized image as input and the algorithm returned a data file that contains a segmented image. Though it went well (99%) for line segmentation with different fonts, it cannot be applied to a complex structure, due to a dependence on the vertical projection in the first phase of page segmentation, where there must always be a vertical white space on the whole page between the articles. Therefore there are incorrectly segmented lines with poor detection in the case of the absence of vertical white space in the page image. In [Abu06] the authors proposed a robust method for line segmentation and they score of 97.8% for both simplified and traditional text fonts (97.3% for simplified font and 98.4% for a traditional font), based on splitting one region into many smaller regions in a repetitive way until no more regions require splitting using a horizontal projection and a set of constraints. However, as the program does not work with a complex structure that has more than one article and different sizes of text on the same page, it is not possible to extract the lines for both normal text or large size text from each article if it exceeds an article on the image page or if it contains variable font size texts in the same article. Another study [Sou10] focused on the line segmentation of low-quality documents, by investigating different text-line segmentation algorithms like Projection Profiles (PP), the Run Length Smearing Algorithm (RLSA) and Adaptive Run-Length Smearing Algorithm (ARLSA); and by applying HPP they achieved a score of 100% on English documents that had varying spaces. RLSA achieved an accuracy of 96% on overlapping documents, and ARLSA achieved an accuracy of 99% on English documents with overlapping components. However, PP cannot handle images where the text lines are overlapping or touching. RLSA and ARLSA fail if there is any overlap between two text lines.

Figure 7: Samples of images used for the line segmentation method: (a) Ibrahim's data [Abu06]: an article with the same text size in one column; (b) Soujanya et al.'s data [Sou10]: an article with a different size font in one column; (c) Ayesh et al.'s data [Mun17]: variability of font size and the possibility of multiple articles, which was restricted by the presence of vertical white spaces between them; (d) Our own data where several articles have different font sizes and figures.

Although the program can handle many size fonts, in the previous study they based it on documents which had just one article hence one title had no more than this, and their approach cannot be applied on pages with a complex structure e.g. when there are many articles, figures, and titles. In [Bou18], they extracted the lines based on horizontal projection, local minima, and conflict resolution and got a score of 99.85% for text. They based it on more than one hundred pages taken from the same newspaper, and they had a similar structure for all the images they had for their dataset. Their text was simple text or titles because they did not distinguish between them, and this made it difficult to compare because a good segmentation line does not mean good title and subtitle detection. In fact, every article has a title, which leads us to think that good article detection with good line segmentation means a good title and subtitle extraction. In their study, they got a score of 90.03% for

ISBN 978-80-86943-38-1

article detection and for title detection they were unable to exceed this even in the best cases.

5 CONCLUSIONS

Title segmentation plays a significant role in the segmentation phase for the identification of any article in any random document paper. In this study, we handled the problem of distinguishing text and overlapping-lines with small font size, and for large fonts, using RLSA, Connected Components (CCs) and Projection Profile (PP) in scanned pages. We evaluated the proposed method on the real three hundred text images using the PATD database. The results presented here are superior to those of existing algorithms that perform the same task. Our future goal is to extend our document language procedure to other document formats and structures and generalize its capabilities.

6 REFERENCES

[Azo95] A. S. Azokly, "A Uniform Approach to the Recognition of the Physical Structure of Composite Documents Based on Spatial Analysis". PhD Thesis, Institute of Computer Science - University of Fribourg (Switzerland), 1995.

[Mon11] F. Montreuil, "Extraction of document structures by conditional random fields: application to handwritten mail processing". PhD thesis, University of Rouen, 2011.

[Gra09] A. Graves and J. Schmidhuber, Offline handwriting recognition with multidimensional recurrent neural networks, in Advances in neural information processing systems, pp.545-552, 2009.

[Ray15] A. Ray, S. Rajeswar, and S. Chaudhury, Text recognition using Deep BLSTM N, in Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on. IEEE, pp.1-6, 2015.

[Ryu14] Ryu, H. I. Koo, and N. I. Cho, Language independent text-line extraction algorithm for handwritten documents, IEEE Signal processing letters, Vol.21, No.9, pp.1115-1119, 2014.

[Buk08] S. S. Bukhari, F. Shafait, and T. M. Breuel, Segmentation of curled text lines using active contours, in Document Analysis Systems, 2008. DAS 08. The Eighth IAPR International Workshop on. IEEE, pp.270-277, 2008.

[Zek11] A. M. Zeki, M. S. Zakaria, and C.-Y. Liong, Segmentation of Arabic characters: A comprehensive survey, International Journal of Technology Diffusion, Vol.2, No.4, pp.48-82, 2011.

[Che01] A. Cheung, M. Bennamoun, and N. W. Bergmann, An Arabic optical character recognition system using recognition based segmentation, Pattern recognition, Vol.34, No.2, pp.215-233, 2001.

[Sou10] P. Soujanya, V. K. Koppula, K. Gaddam, and P. Sruthi, Comparative study of text line segmentation algorithmson low quality documents, CMR College of Engineering and Technology Cognizant Technologies, Hyderabad, India, 2010.

[Hus15] S. Hussain, S. Ali et al., Nastalique segmentation-based approach for Urdu OCR, International Journal on Document Analysis and Recognition (IJDAR), Vol.18, No.4, pp.357-374, 2015.

[Alj12] I. Aljarrah, O. Al-Khaleel, K. Mhaidat, M. Alrefai, A. Alzu bi, M. Rababah et al., Automated system for Arabic optical character recognition with lookup dictionary, Journal of Emerging Technologies in Web Intelligence, Vol.4, No.4, pp.362-370, 2012.

[Bro13] D. Brodic and Z. N. Milivojevi, Text line segmentation with the algorithm based on the oriented anisotropic Gaussian kernel, Journal of Electrical Engineering, Vol.64, No.4, pp.238-243, 2013.

[Bou18] H.Bouressace and J.Csirik, Recognition of the logical structure of Arabic newspaper pages, 21st International Conference on Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence (Springer), Vol.11107, No.3, pp.251-258, 2018.

[Mun17] Muna Ayesh, Khader Mohammad, and Aziz Qaroush, A Robust Line Segmentation Algorithm for Arabic Printed Text with Diacritics, IS and T International Symposium on Electronic Imaging, pp.42-47, 2017.

[Zek11] A. M. Zeki, M. S. Zakaria, and C.Y. Liong, Segmentation of Arabic characters: A comprehensive survey, International Journal of Technology Diffusion, Vol.2, No.4, pp.48-82, 2011.

[Gor97] L. O Gorman and R. Kasturi, Executive briefing: document image analysis, IEEE Computer Society Press, ISBN 0-8186-7802-X, 1997.

[Ari07] M. Arivazhagan, H. Srinivasan, and S. Srihari, A statistical approach to line segmentation in handwritten documents, International Society for Optics and Photonics, 2007.

[Abu06] I. Abuhaiba, Segmentation of discrete Arabic script document images, Al Azhar University, Vol.8, No.1810-6366, pp.85-108, 2006.

[Sim08] Simone Marinai, Introduction to Document Analysis and Recognition, Studies in Computational Intelligence (SCI), Vol.90, No.1-20, 2008.

[Bou19] H.Bouressace and J.Csirik, Printed Arabic Text Database for Automatic Recognition Systems, 5th International Conference on Computer and Technology Applications, pp.107-111, 2019.

 ISBN 978-80-86943-38-1