ISSN 2278-3091
Volume 8, No.1.3, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse2181.32019.pdf
https://doi.org/10.30534/ijatcse/2019/2181.32019

Frame Removal For Mushaf Al-Quran Using Irregular Binary Region

Laith Nazeeh Bany Melhem[1], Mohd Sanusi Azmi[1], Nur Atikah Arbain[1], Azah Kamilah Muda[1], Intan Ermahani A. Jalil[1], Jabril Ramdan[2],
[1]Faculty of Information and Communication Technology (FTMK), Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia, lnm1989@gmail.com
[2]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43200 Bangi, Selangor, Malaysia

ABSTRACT

Segmentation is a process to remove frame or frame exists in each page of some releases of mushaf Al-Quran. The fault in segmentation process affects the holiness of Al-Quran. The difficulty to identify the appearance of frame around text areas as well as noisy black stripes has caused the segmentation process to be improperly carried out. In this paper, an algorithm for detecting the frame on Al-Quran page without affecting its content is proposed. Firstly, preprocessing was carried out by using the binarisation method. Then, it was followed with the process of detecting the frame in each page. In this stage, the proposed algorithm was applied by calculating the percentage of black pixel of binary from vertical (column) to horizontal (row). The results, based on experiments on several Al-Quran pages from different Al-Quran styles, demonstrate the effectiveness of the proposed technique.

Key words : Al-Quran, Binarisation, Detection, Frame, Segmentation.

1. INTRODUCTION

Image processing is a popular research area in computer science. Today, image processing is not only focused on fundamental issues addressed by researches, but it also addresses the suitability of the research into several domains, such as biometric[1], character recognition [2] and document analysis [8].

Besides, Khairuddin Omar (2010) and Mohammad Faidzul et al. (2010) mentioned that the image processing body knowledge consists of several phases starting from data collection, preprocessing, feature extraction, feature selection, classification, and post processing [2], [9]-[10]. Each phase in the image processing has sub-processes. For example, the preprocessing phase for the Arabic/Jawi character recognition is categorised into Binarisation, Edge Detection, Thinning, and Segmentation before Feature Extraction takes place [2].
In this paper, the focus is on the segmentation for removing frame exist in mushaf holy Quran. Thus, the segmentation for removing frame exists in mushaf holy Quran is based on the processes for segmenting Arabic/Jawi handwritten texts.

Holy Quran is the book of Allah SWT. Al-Quran consists of 30 chapters, 114 surahs and 6236 verses. However, the number of pages and lines are different based on the publishers.

Arabic language (i.e. the language in which the Al Quran is written) has markings called “diacritical marks” or “diacritics”, which represent short vowels or other sounds, and if one of these diacritical marks is ignored it will change the meaning of the word. There are many Arabic character recognition techniques, which can recognise the characters of the text or the whole page (Khairuddin Omar, 2000; Mohamad Faidzul, 2010) but all of these techniques provide recognition for characters without considering the diacritical marks, which may affect the meaning of the Al Quran’s word and the holiness of Al-Quran.

Segmentation process for segmenting Al-Quran needs to be studied carefully. This is because Al-Quran is the book of Allah SWT. Any incorrect segmentation will affect the holiness of Al-Quran.
In this paper, one technique for segmenting Al-Quran is proposed. The technique considers diacritical marks (Tashkil) in order to protect the holiness of Al-Quran.

2. RELATED WORK

There are many research on segmentation, and segmenting the frame of mushaf Al-Quran needs to be done carefully in order to preserve the holiness of Al-Quran. Studies, which are near to Al-Quran segmentation, are Arabic and Jawi segmentations, but, both are not suitable to be applied for Al-Quran due to diacritical marks (Tashkil), words, and sentences.

The first step before the segmentation phase is preprocessing the documents to produce a clean image of the documents. Mushaf Al-Quran page image contains frame. Detecting and removing these unwanted areas is critical in order to achieve better text segmentation results. Before the frame detection and removal take place, it is processed into an image binarisation using Otsu method [11].

There are some segmentation techniques for Arabic and Jawi characters [2]. One of the techniques has categorised the preprocessing phase for the Arabic/Jawi character recognition

into Binarisation, Edge Detection, Thinning, and Segmentation before Feature Extraction takes place [2]-[6]. In [7], they used the triangle features to the feature extraction method for better classification.

2.1. Frame removal

The most common approach to eliminate marginal noise is by performing document cleaning by filtering out connected components based on their size and aspect ratio. However, when characters from the adjacent page are also present, they usually cannot be filtered out using only these features. There are only a few techniques in the bibliography for page borders detection, and they are mainly focused on printed document images.

Le et al. [12] proposed a technique for removing the border, which is predicated on the classification of blank, non-textual and textual columns and rows, the border objects location, and an analysis on crossing counts of textual squares and projection profiles. There are many uses of this approach. Moreover, it is assumed that the page borders are very close to the image edges and the separation of the borders from the contents of the image is by a blank space. However, this assumption is often violated. Fan et al. [13] proposed a technique for detecting the black noisy regions that overlap with the text, but the scholars did not assume there were noisy text regions. They proposed a framework to reduce the image resolution in order to detect and remove the black borders, which hides text, by threshold filter and thus, leaving the border of the image. They applied the deletion process on the original image.

In [14], Avila et al. proposed algorithms for non-invading and invading borders that work as "flood fill" algorithm. The algorithm for the "non-invading border" supposes that the information of the document merges with the border of noisy black. In the connected area, it uses two parameters that are related to the document, which are the segment maximum size that belongs to the document text, and the maximum distance between the lines in order to curb the flooding. On the other hand, the algorithm for the "invading" supposes that the black areas are invaded by the borders of noisy black. If the border of noisy black merges with the text region of the document, all areas and the part of the text region are removed and flooded.

2.2. Image segmentation

In order to extract features from the text image, it should be segmented into lines, words, characters or primitives. Arabic OCR systems are classified into two major types, depending on the employed method of segmentation (i.e., segmentation based systems and segmentation free systems). The segmentation procedure is a major challenging phase for any Arabic OCR system because of the cursive nature of the Arabic script. This challenge occurs at segmentation based systems while segmentation free systems avoid this problem.

Various document image segmentation techniques were proposed in the literature. These techniques can be categorised based on the document image segmentation algorithm that they adopt. Among the renowned segmentation algorithms are: X–Y cuts or projection profiles based [15], Run Length Smoothing Algorithm (RLSA) [16], component grouping [17], document spectrum [18], whitespace analysis [19], constrained text lines [20], Hough transform [21]-[22], Voronoi tessellation [23], and Scale space analysis [24]. All of the aforementioned segmentation algorithms are mainly designed for contemporary documents.

3. RESEARCH FRAMEWORK

The conceptual framework of this paper is divided into two phases. The first section is the investigation phase, included the background of the research, problems, current issues of the domains that determine the scope of the research, and the aim of the research. And the second section is the implementation phase.

4. EXPERIMENTAL SETUP

In this paper, several experimental tests were conducted to seek out the most effective practice techniques for segmenting images of holy Quran pages. The experiment used is predicated on the best percentage to detect the frame, and to segment Al-Quran pages into pages without frame. The experiment test is explained along with the objectives. The algorithmic programmes used, the input and the results obtained from the algorithm are also disclosed.

- Experiment I
  i. Objective: To get the best percentage to detect the frame, and to segment Al-Quran pages into pages without frame
  ii. Input: Data as images from Al-Quran pages
  iii. Algorithm: The proposed method
  iv. Output: Image of each page without frame

5. IMPLEMENTATION PROCESS

Fig. 1 presents a framework of techniques, which enables page segmentation of a set of mushaf Al-Quran. In this paper, there are four classification of methods in segmenting the mushaf Quran pages into pages without frame, which are: (a) The preprocessing, which encompass binarisation, and noise removal is applied; (b) The frame on the pages are detected; (c) Segments to page without frame; and (d) The segmented pages are saved as image.

The following are detailed explanations of all the steps used in the process of mushaf Al-Quran pages segmentation.

5.1. Image Binarisation

In the preprocessing step of the document analysis, the Binarisation was performed, and it was designed to segment the text from the document background. There are many

algorithms that have been proposed to do the Binarisation task of a document. In this paper, the Otsu method was used to do the Binarisation [11].

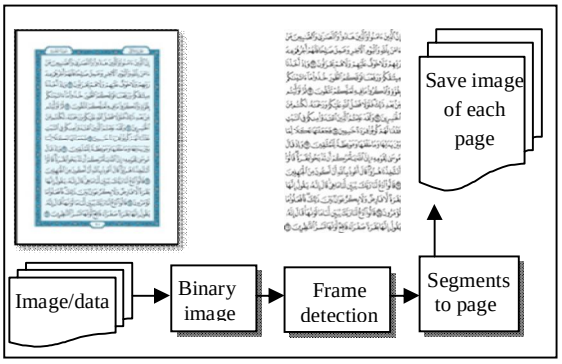

Figure 1: The Frame Removal in Al-Quran and Segmentation to Pages

5.2 Page Segmentation

In this study, blank space and frame were detected and removed from the mushaf holy Quran pages by the suggested methodology. The blank space and frame removal method relies upon the density of the binary values (the binary representation).

In Figure 2, a new methodology was proposed to detect the frames of page by three frames, which are: i) The exterior blank space and marginal frame; ii) Frame; and iii) The interior blank space. That was based on the horizontal and vertical white pixel percentages. This study aimed to segment Al-Quran page image into page without blank space or frame.

i. Frame 1 (The exterior blank space and marginal frame) :
In the first frame step, the blank space and marginal frame of the page were removed. In order to achieve this, it was first processed into an image processing and converted into a binary representation to calculate the total frequency of zero value (white colour).

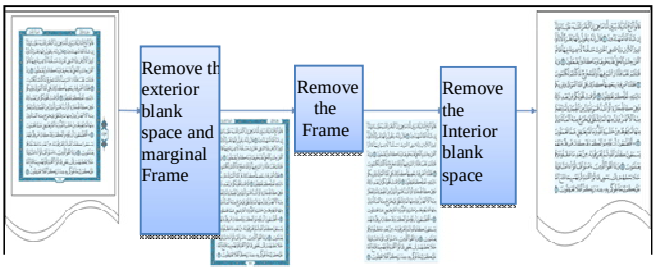

Figure 2: The Process of Removing the Exterior Blank Space and Marginal Frame from Al-Quran Page

Consider the input page grayscale image with the dimension of X × Y. The paper aimed to find the frames of the page defined by the new coordinates as demonstrated in Fig. 2. The constant value is assumed as: White Percentage (WPer) = White Percentage from Out Frame (WPerout) = 45.0, for this frame. In the next step, which includes the searching of horizontal detection of the first frame edges, the White Percentage from Up (WPerU) for each line starting from line 0 to Y-1 was calculated by using the following formula:

$$WPerU_l = \frac{100 \sum_{i=0}^{X} \mathrm{WPix_l}}{X}, \text{where } 0 \le l < Y, \quad WPix: White\ pixel \quad (1)$$

When the value for WPerUl < WPer, the process of detecting the upper limit would stop and it was identified as the upper limit. Then, the White Percentage was calculated from Down (WPerD) for each line, starting from Y-1 to 0, by using the following formula:

$$WPerD_l = \frac{100 \sum_{i=0}^{X} \mathrm{WPix_l}}{X}, \text{where } Y < l \le 0, \quad WPix: White\ pixel \quad (2)$$

When the value for WPerDl < WPer, the process of detecting the down limit would stop and it would identify l+1 as the down limit. In the next step, which includes the searching of vertical detection of the first frame edges, the White Percentage from left (WPerL) for each line starting from line 0 to X-1 was calculated by using the following formula:

$$WPerL_l = \frac{100 \sum_{i=0}^{Y} \mathrm{WPix_l}}{Y}, \text{where } 0 \le l < X, \quad WPix: White\ pixel \quad (3)$$

When the value for WPerLl < WPer, the process of detecting the left limit would stop and it would identify l as the left limit. Then, the White Percentage was calculated from right (WPerR) for each line, starting from line X-1 to 0 by using the following formula:

$$WPerR_l = \frac{100 \sum_{i=0}^{Y} \mathrm{WPix_l}}{Y}, \text{where } X < l \le 0, \quad WPix: White\ pixel \quad (4)$$

When the value for WPerRl < WPer, the process of detecting the right limit would stop and it would identify l+1 as the right limit. Frames were detected after detecting the horizontal and vertical frame edges. Once the frame was detected, a new image was cropped at the point of (left, upper) and with the size of right-left × down-upper. The next page frame was identified and the frame was generated using a cropping operation, as shown in Figure 2.

ii. Frame 2 (Frame):
In the second frame step, the frame of the page was removed. In order to achieve this, the result of the previous frame was proceeded to calculate the total frequency of zero value for the

new image by considering the input of new image with the new dimension of X × Y. This study aimed to find the next frame of the page defined by using the new coordinates as demonstrated in Figure 3.

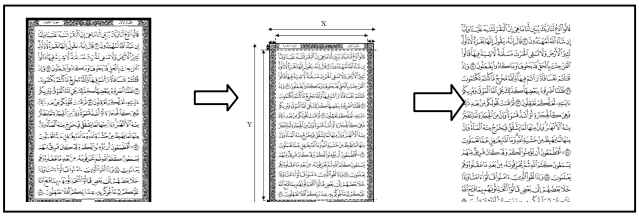

Figure 3: The Process of Removing the Frame from Al-Quran Page

The constant value was assumed as: WPer = White Percentage in Frame (WPeri) = 80.0, for this frame. The next step includes the searching of horizontal detection of the second frame edges by calculating the value of (WPerU) for each line, starting from line 0 to Y-1 by using Formula (1). When the value for WPerUl > WPer, for five consecutive rows, the process of detecting the upper limit would stop and it identified l as the upper limit.

Then, the (WPerD) was calculated for each line, starting from Y-1 to 0 by using Formula (2). When the value for WPerDl > WPer for five consecutive rows, the process of detecting the down limit would stop and it identified l as the down limit. The next step includes the searching of vertical detection of the first frame edges by calculating the (WPerL) for each line, starting from line 0 to X-1 by using Formula (3). When the value for WPerLl > WPer for five consecutive rows, the process of detecting the left limit would stop and it identified l as the left limit.

Then, the value of (WPerR) was calculated for each line, starting from line X-1 to 0 by using Formula (4). When the value for WPerRl > WPer for five consecutive rows, the process of detecting the right limit would stop and it identified l as the right limit. Frames were detected after detecting the horizontal and vertical frame edges. Once the frame was detected, a new image was cropped at the point of (left, upper) and with the size right-left × down-upper. The next page frame was identified and frame was generated using a cropping operation, as shown in Figure 3.

iii. Frame 3 (The interior blank space):
In the third frame step, the interior blank space of the page was removed. In order to achieve this, the result of the previous frame was firstly proceeded to calculate the total frequency of zero value for the new image by considering the input of the new image with the new dimension of X × Y. This study aimed to find the next frame of the page defined by the new coordinates as demonstrated in Fig. 3.
The constant value was assumed as: WPer = White Percentage from interior blank space (WPerin) = 100.0, for this frame. The next step includes the searching of horizontal detection of the second frame edges by calculating the value of (WPerU) for each line, starting from line 0 to Y-1 by using Formula (1).
When the value for WPerUl < WPer, the process of detecting the upper limit would stop and it identified l as the upper limit. Then, the value of (WPerD) was calculated for each line, starting from Y-1 to 0 by using Formula (2).

When the value for WPerDl < WPer, the process of detecting the down limit would stop and it identified l+1 as the down limit. The next step includes the searching of vertical detection of the first frame edges by calculating the value of (WPerL) for each line, starting from line 0 to X-1 by using Formula (3).
When the value for WPerLl < WPer, the process of detecting the left limit would stop and it identified l as the left limit. Then, the value of (WPerR) was calculated for each line, starting from line X-1 to 0 by using Formula (4). When the value for WPerRl < WPer, the process of detecting the right limit would stop and it identified l+1 as the right limit.

Frames were detected after detecting the horizontal and vertical frame edges. Once the frame was detected, a new image was cropped at the point of (left, upper) with the size of right-left × down-upper. Al-Quran page was identified and the page without frame was generated using a cropping operation, as shown in Figure 4.

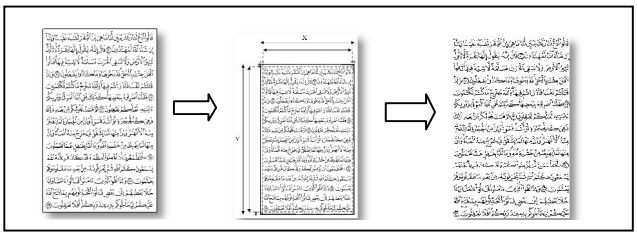

Figure 4: The Process of Removing the Interior Blank Space from Al-Quran Page

5. TESTING AND RESULT

Testing occurred throughout the various stages of the application development so as to make sure an adequate performance during page segmentation. The testing includes checks for both, the word and the diacritical marks. With concern to the correctness of the software, informal tests were applied in every completed page. This especially dealt with Al-Quran pages segmentation as to make sure that all of the pages were segmented and tested for validity, and did not contain any missing word or diacritical marks.

A lot of experiments were done to verify the validity of the proposed methodology and for each experiment the segmentation percentage was changed to suit all of the Al-Quran writing styles. Different styles were used for different Al-Quran writing styles and different pages. In order to calculate the precision in segmentation, the density of

frame was calculated for each page frame in all Al-Quran writing styles and then the average values of them were extracted.

After several tests, it is concluded that the following results are the best results obtained and desired, without missing any of the text or the diacritical marks from Al-Quran pages. Fig. 5 illustrates the results of a sample page after applying page segmentation in which the frame was removed.

From the tests and results, it is shown that the segmentation process for several copies of Al-Quran is a success and the images of the pages can be saved without frame. That means the possibility of using the proposed methodology to segment the Al-Quran pages without any missing word or diacritical marks. Therefore, the segmentation process preserves the holiness of Al-Quran.

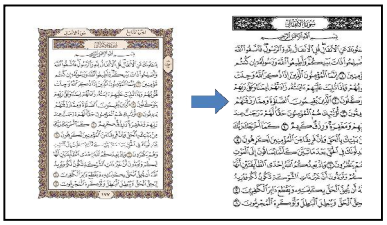

Figure 5: The Result of Page Segmentation

CONCLUSION

The Holy Quran is very important to the Muslims with respect to its authenticity. In this paper, a novel removal frame technique from the holy Quran pages was proposed, which enables us to remove the frame of the page according to the percentage of the binary numbers in the Quran page image. Any incorrect segmentation will affect the holiness of Al-Quran. Subsequently, this study aimed to segment Al-Quran pages to pages without frame, without any changes on the content. First, a preprocessing was applied, which includes binarisation. Then, the frame of Al-Quran pages was detected. In this stage, the vertical and horizontal white percentages were introduced, which have been proven efficient for detecting the frame. The results are based on several Al-Quran pages from different Al-Quran styles in order to demonstrate the effectiveness of the proposed technique.

ACKNOWLEDGEMENT

The authors would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for funding PJP/2016/PBPI/HI4/S01476, also to the Ministry of Higher Education to fund my study under Malaysia International Scholarship (MIS). The authors would also like to extend their thanks the Faculty of Information Communication and Technology, UTeM for providing excellent research faculties and facilities.

REFERENCES

1. Phillips, P. J., McCabe, R. M., & Chellappa, R. **Biometric image processing and recognition.** Signal Processing Conference (1998):1-8.
2. Omar, K. **Pengecaman Tulisan Tangan Teks Jawi Menggunakan Penkelas Multiaras**, Univ. Putra Malaysia. (2000).
3. M. S. Azmi, K. Omar, M. Faidzul, N. Khadijah, and W. Mohd, **Arabic Calligraphy Identification for Digital Jawi Paleography using Triangle Blocks**. no. July. (2011). https://doi.org/10.1109/ICEEI.2011.6021785
4. M. S. Azmi, K. Omar, M. F. Nasrudin, B. Idrus, and K. Wan Mohd Ghazali. **Digit recognition for Arabic/Jawi and Roman using features from triangle geometry**, AIP Conf. Proc. 1522 (2013) 526–537. https://doi.org/10.1063/1.4801171
5. M. S. Azmi and K. Omar. **Features Extraction of Arabic Calligraphy using extended Triangle Model for Digital Jawi Paleography Analysis**, Int. J. Comput. Inf. Syst. Ind. Manag. Appl. 5 (2013) 696–703.
6. M. S. Azmi, M. F. Nasrudin, K. Omar, C. W. S. B. C. W. Ahmad, and K. W. M. Ghazali**. Exploiting features from triangle geometry for digit recognition**, 2013 Int. Conf. Control. Decis. Inf. Technol. CoDIT. (2013) 876–880. https://doi.org/10.1109/CoDIT.2013.6689658
7. N. A. Arbain, M. S. Azmi, L. B. Melhem, A. K. Muda, and H. Rashaideh. **Enhancement of Triangle Coordinates For Triangle Features For Better Classification**, Jordanian J. Comput. Inf. Technol. 2(2) (2016) 108–119. https://doi.org/10.5455/jjcit.71-1448511760
8. Sauvola, J., and Pietikäinen, M. **Adaptive document image binarization, Pattern Recognition**. 33(2) (2000) 225–236. https://doi.org/10.1016/S0031-3203(99)00055-2
9. Azmi, M. S., Omar, K., Nasrudin, M. F., & Muda, A. K. **Fitur Baharu Dari Kombinasi Geometri Segitiga dan Pengezonan utk Paleografi Jawi Digital** (2013).
10. Nasrudin, M. F., Omar, K., Liong, C. Y., & Zakaria, M. S. **Pengecaman aksara jawi menggunakan jelmaan surih, Sains Malaysiana**. 39(2) (2010) 291–297.
11. Otsu, N. A Tlreshold **Selection Method from Gray-Level Histograms.** Automatica. 11(285-296) (1975) 23-27.
12. Le, D. X., Thoma, G. R., & Wechsler, H. **Automated borders detection and adaptive segmentation for binary document images.** Proc. - Int. Conf. Pattern Recognit. 3(1996) 737–741. https://doi.org/10.1109/ICPR.1996.547266
13. Fan, K. C., Wang, Y. K., & Lay, T. R. **Marginal noise removal of document images. Pattern Recognit**. 35(11) (2002) 2593–2611. https://doi.org/10.1016/S0031-3203(01)00205-9
14. Ávila, B. T., & Lins, R. D. **A new algorithm for removing noisy borders from monochromatic documents**. Proc. 2004 ACM Symp. Appl. Comput. (2004) 1219-1225.

https://doi.org/10.1145/967900.968149
15. Nagy, G., & Seth, S. **Hierarchical representation of optically scanned documents**. Proc. Int. Conf. Pattern Recognit. 1(1984) 347–349.
16. Wahl, F. M., Wong, K. Y., & Casey, R. G. **Block segmentation and text extraction in mixed text/image documents**. Computer Graphics and Image Processing. 20(4) (1982) 375-390. https://doi.org/10.1016/0146-664X(82)90059-4
17. Feldbach, M., & Tonnies, K. D. **Line detection and segmentation in historical church registers**. Proc. Sixth Int. Conf. Doc. Anal. Recognit. (2001) 743-747. https://doi.org/10.1007/3-540-45404-7_19
18. O'Gorman, L. **The Document spectrum for page layout analysis.** IEEE Trans. Pattern Anal. Mach. Intell. 15(11) (1993) 1162–1173. https://doi.org/10.1109/34.244677
19. Baird, H. S. **Background structure in document images. International Journal of Pattern Recognition and Artificial Intelligence**. 8(05) (1994) 1013–1030. https://doi.org/10.1142/S0218001494000516
20. Breuel, T. M. **Two Algorithms for Geometric Layout Analysis**. In Proceedings of the Workshop on Document Analysis Systems, Princeton. NJ, USA. (2002) 188–199. https://doi.org/10.1007/3-540-45869-7_23
21. Hough, P. V. **Method and means for recognizing complex patterns** (No. US 3069654). (1962).
22. Duda, R. O., & Hart, P. E. **Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM**. 15(1) (1972) 11–15. https://doi.org/10.1145/361237.361242
23. Kise, K., Sato, A., & Iwata, M. **Segmentation of Page Images Using the Area Voronoi Diagram**. Comput. Vis. Image Underst.70(3) (1998) 370–382. https://doi.org/10.1006/cviu.1998.0684
24. Manmatha, R., & Rothfeder, J. L**. A scale space approach for automatically segmenting words from historical handwritten documents**. IEEE Trans. Pattern Anal. Mach. Intell. 27(8) (2005) 1212–122. https://doi.org/10.1109/TPAMI.2005.150