Indonesian Journal of Electrical Engineering and Computer Science
Vol. 18, No. 1, April 2020, pp. 478~484
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v18.i1.pp478-484

Diacritic segmentation technique for arabic handwritten using region-based

Ahmed Abdalla Sheikh, Mohd Sanusi Azmi, Maslita Abd Aziz, Mohammed Nasser Al-Mhiqani, Salem Saleh Bafjaish
Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi,
Unvirsiti Teknikal Malaysia Melaka, Malaysia

Article Info

Article history:

Received Aug 11, 2019
Revised Oct 19, 2019
Accepted Nov 6, 2019

Keywords:

Arabic handwritten segmentation
Arabic diacritics
Diacritics segmentation
Region-Based

ABSTRACT

Arabic is a broadly utilized alphabetic composition framework on the planet, and it has 28 essential letters. The letters in order was first used to compose messages in Arabic, most prominently the Qur'an the holy book of Islam. However, Arabic language has diacritics in the word or letters which are not something extra or discretionary to the language, rather they are a vital piece of it. By changing some diacritics may change both the syntax and semantics of a word by turning a word into another. However, the current researches address the foreground image and consider the diacritics as noises or secondary images. Thus, it is not suitable for Arabic handwritten. The diacritics will be removed from the image and this will lead to losing some good features. Furthermore, to extract the diacritics, the region-based segmentation technique is used. The image will be measured based on the region properties by first finding the connected component in binary image, and then we will determine the best area range measurement in that region for each image. The proposed technique region based has been tested in nine different images with different handwritten style, and successfully extracted secondary foreground images (diacritics) for each image

Corresponding Author:

Ahmed Abdalla Sheikh,
Department of Information and Communication Technology,
Unvirsiti Teknikal Malaysia Melaka,
Jalan Hang Tuah Jaya, 76100 Durian Tunggal, Melaka.
Email: ashiekh295@gmail.com

1. INTRODUCTION

Arabic language is an unvirsal around the world, and it is an official language for 25 nations of populace over than 250 million and it consist both letters and diacritics [1]. In any case, Arabic composition construction has fundamental contrasts contrasted with Latin and East Asia. Arabic has 28 letters [2] and second most commonly used alphabetic writing system [3]. A few letters are indistinguishable in frame and have optional item to separate it. Arabic characters are written from right to left (RTL) cursively [4]. Each letter changes its frame contingent upon its situation in the word. Moreover, [5] writing of Arabic script is normally under indicated for short and long vowels [6] and other markup, alluded to as diacritics. Besides, the Arabic writing has numbers of diacritics, including i'jam ⟨إِعْجَام⟩, and tashkil ⟨تَشْكِيل⟩. The latter include the ḥarakāt ⟨حَرَكَات⟩. Some diacritics in Arabic writing are optional to represent the missing vowels and consonant length. For some Arabic letters they have dots, these dots can be put above or underneath the letter body, and dabs can be single, twofold or triple. Diacritics restoration is one of the real difficulties for the Arabic common language processing. In fact, the nonappearance of vowels in Arabic writings creates a significant number of ambiguities in morphological, syntactic and semantic layer [7]. in other hand, [8] the diacritics restoration in the text will simplify its pronunciation and proper understanding. Thus, the Arabic diacritic is not something extra or an optional to the language, rather it considers as a vital part of the language. Moreover, for mastering

the format of writing style of the Arabic alphabet, the author or the writer will have to to master the correct placing of the diacritics and as well as the writing.

Moreover, Arabic is an extensively used alphabetic synthesis system on the planet, and it has 28 fundamental letters. The letters all together was first used to make messages in Arabic, most unmistakably the Qur'an the heavenly book of Allah swt. With spread of Islam it came to be used form various vernaculars like Urdu, Pashto, Uyghur (in China), Ottoman Turkish and Spanish (in Western Europe). By then various changes and enhancements have been made to Arabic composed work substance, which understands extra letters and strokes [9]. The new strokes are called diacritics, as shown in Figure 1, and the clarification behind adding these diacritics is to see. In Arabic language especially in (Quran Language) there's signs called by "diacritics marks" (Tashkil) [10] which represent short vowels or other sounds if one of these diacritical marks overlooked they might change the whole meaning of the word [11]. The [12] usage of diacritics in Arabic language is an optional, not very common, practice in modern standard Arabic, except for holy Al Quran.

Connectivity is an outcome of the Arabic cursive nature script. Arabic text can only be scripted cursively [13]. However, 8 out of the 28 Arabic letters do not connect to subsequent letters. In addition, even connectable letters do not connect to subsequent letters when the end of the word has been touched. Arabic language [14] utilizes diverse marks to distinguish between words that have same letters. It relies upon Arabic Diacritics (Harakat), as its known in Arabic [15] where Diacritics are discretionary. The majority of Arabic contents can be read without Diacritics which relies upon language syntax.

In addition, Diacritical marks play a crucial role in meeting the criteria of ease of use of typographic text, for example, homogeneity, lucidity and neatness. To change the diacritic of a letter in a word could totally change its semantic or meaning [16]. The circumstance is extremely entangled with multilingual content. To be sure, the issue of configuration turns out to be progressively troublesome by the nearness of diacritics that originate from different contents; they are utilized for various purposes and are controlled by different typographic tenets. It is quite challenging to adapt rules from one script to another [17].

The Figure 2 shows and proves the difficulties and challenges of Arabic diacritics, as we can see in the figure there are four similar words but having different diacritics marks each. The first word which is (عَلَمْ) which is stands for word (Flag). The second word which is word (عَلَّمْ) which stands for word (Taught). The third word which is (عِلْمْ) which stands for word (Science). And finally, the fourth word which is (عَلِمْ) which is stands for word (Knew). All these words have same writing style, but they differ in terms of diacritical marks, if one of these words got any missed placed diacritics mark definitely the whole meaning will turn out which will increase the ambiguousness of the word.

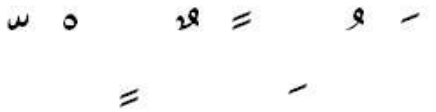


Figure 1. Arabic diaciritcs marks [12]

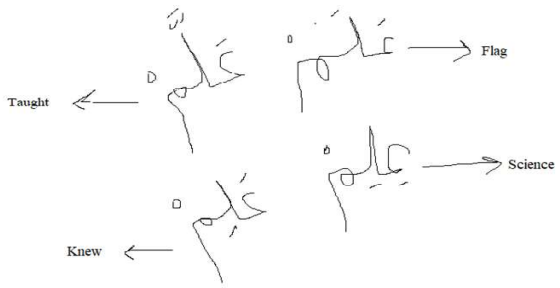


Figure 2. Different diacritical marks placement in same word

This is the main challenges in Arabic handwritten document how to recognize the meaning of the word when there is no diacritics placed in the word, it will be ambiguous and hard to recognize for non-natives Arabs and as well as it could affect the originality of document of Arabic language that contains many diacritics marks [18], if any missing or missed placed diacritics it will affect its holiness and as well as its originality. However, in this research the segmentation process will be used to prepare image for extraction process. Arabic handwritten document will be processed from the image in order to segment the diacritics efficiently from the document. Thus, based on the extracted document from the image, our segmentation method will be able to detect the secondary foreground image (diacritics) that presented in the document. The proposed technique is region-based segmentation which is the simplest technique for extracting and segmenting the secondary foreground (diacritics) from Arabic handwritten image.

2. MATERIAL AND METHOD

In this paper, the data has been collected from Arabic handwritten images and as well Al Quran since al Quran originally is a handwritten document. There are nine different images seven of them are written by the author while the other two images been collected from Mus'haf al-Madinah printing complex [19]. Besides, Arabic has different style of handwriting in both shape and size [20] and as well as it contains many diacritics. The diacritic here is referring to dots and marks in every word of Arabic language. The following Figure. 3 shows the overall steps and method for extracting diacritics from image.

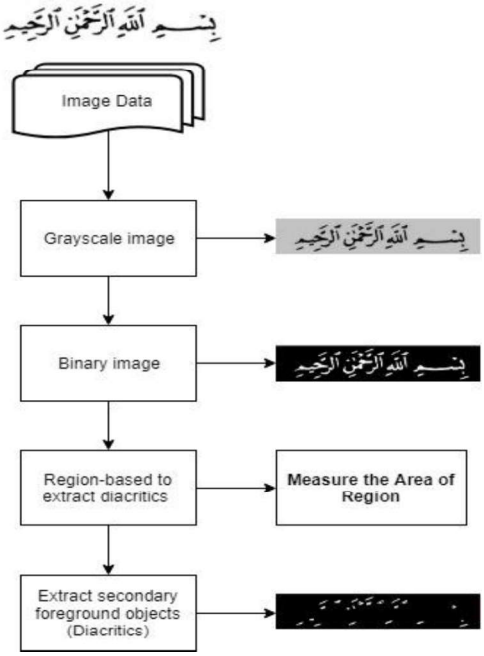


Figure 3. Method for extracting diacritics

2.1. Image Data

The image data in Figure 4 has been collected from Arabic handwritten images and as well Al Quran since al Quran originally is a handwritten document. There are different images. Besides, Arabic has variability style of writing [21] and as well as it contains many diacritics. The diacritic here is referring to dots and marks in every word of Arabic language.

2.2. Grayscale

Initially, it's clear that some of images come with different colors, therefore there is a need for the conversion to grayscale image in order to make the process of extracting secondary foreground easier. Mushaf Al-Quran images or Arabic handwritten document must be converted to grayscale in Figure 5 in order to make it easy later on once converting to binary image.

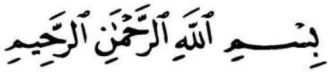


Figure 4. Image data

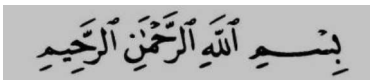


Figure 5. Grayscale image

2.3. Binary

After the image converted to grayscale form, another pre-processing step is connected on the image which is the Binarization. Binary images in Figure 6 are the images that have two values for every pixel, the two conceivable qualities are black and white [22]. Be that as it may, in this stage a binary image is made from the first image so as to help to detect only the significant areas [13] and parts of Arabic handwritten images.

2.4. Region-based to extract diacritics

In this research we have chosen region-based segmentation as a technique for extracting the secondary foreground (diacritics) from Arabic handwritten image. The fundamental objective of segmentation is to segment an image into areas or regions. Some segmentation techniques, for example, thresholding accomplish this objective by searching for the limits between districts dependent on discontinuities in grayscale or shading properties. Regions in an image are a collection of joint pixels with comparable propertie [23], region based segmentation is a technique for deciding the district straightforwardly. The methodology for implementing region-based segmentation technique for extracting secondary foreground image are as following:

a) Use function (bwconncomp) in order to find the connected component in binary image.
b) Use function (regionprops) to measure the properties of image regions, in other hand, it measures the set of properties for each connected component in binary image, and it can be used in contagious regions and dissentious regions.
c) Use function LabelMatrix to create LabelMatrix for function L, from the connected component that returned by the image, and it returns its label matrix in the smallest numeric class necessary for the number of objects.
d) Use function is member because it's helpful with regionprops, bwconncomp, and labelmatrix for making a paired image containing just items or areas that meet certain. For instance, these directions make a binary image containing just the locales whose zone is more prominent than 200 and whose unusualness is under 280. And, it used to compute the desired binary image [24].

2.5. Extract diaciritcs from image

In this section, after we implemented the methodology of region-based segmentation of, the secondary foreground (diacritics) extraction will take place to obtain the image output that contains only diacritics (in Figure 7), and thus the research requirement will be fulfilled.

Figure 6. Binary image

Figure 7. Extracted diacritics

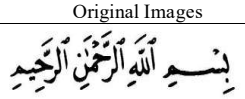
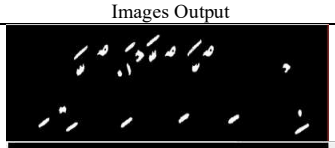
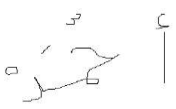
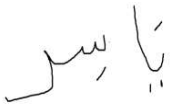
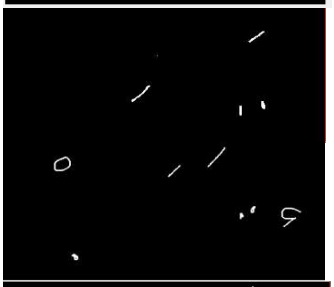
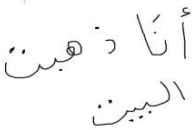
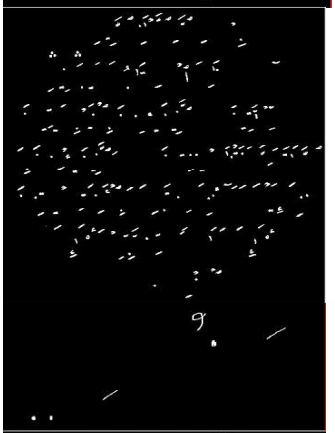
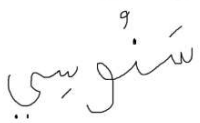
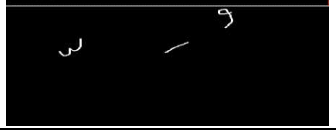
3. RESULTS AND DISCUSSION

In this paper, the proposed technique was tested in Arabic handwritten document including Al Quran. In addition, 10 images of Arabic handwritten were tested by our proposed technique. The proposed technique was implemented using MATLAB software. However, based on the Table 1 there are nine different images with different handwritten style each one of them has their own area measurement due to the differences of the handwritten style and also each image has different pixels from one another. We have to determine the best range for each image in order to achieve the objective which is extracting secondary foreground image (diacritic). For instance, if the area measurement range for the nine images are less than the specified area range in Table 1, there will be some missing secondary foreground objects (diacritics) which means this area is not the desired area. Meanwhile, if the area measurement range for the nine images are greater than the area we specified in Table 1, then it means that this area range is too high, and it will extract the primary foreground objects (huruf) which is not the desired area. Besides, there are two images which are (6,7) in Table 1 that contains missing secondary foreground objects, although the area range for the two images (6,7) are the best area range but the they have many overlapped and connected components in the region which is difficult to recognize between the primary foreground and secondary foreground objects. Therefore, from the area measurements we have conducted, we have noticed that each image has its own area of measurement and its nearly difficult to estimate the best area measurement that can be used for each image to extract the secondary foreground objects from images due to differences of handwritten style, overlapped and several connected components for each image and as well as the different pixels from one image another.

As already known, [25] image segmentation is most basic system of image processor. Notwithstanding, the primary objective of segmentation is to segment image into districts, region based segmentation is a system that deciding the area specifically, and furthermore is the least difficult segmentation method that we could use in our research that centers around extricating diacritics from images. The methodology of implementing region-based segmentation is first to find the connected component in binary image, and then we measure the properties of image regions, in other hand, it measures the set of properties for each connected component in binary image, since diacritics are very small sized objects and determining

the thresholding area region properties would be based on low area measurements, by means that the lower area in the region the smaller objects we obtain which is diacritics, and the higher area in the region the higher objects we obtain which is primary objects. As we can see the concept of implementing region-based segmentation in our research is very simple and solved the research problem efficiently, comparing to other existing techniques that may have different and long implementation process than the proposed technique region-based.

Table 1. Sample Results for Extracting Diacritics Using Region-Based

No	Original Images	Best Area Measurement to Extract Diacritics	Images Output
1		Area >=240 && area <= 279	
2		Area >=55 && area <=80	
3		Area >=94 && area <=525	
4		Area >=282 && area <=293	
5		Area >=280 && area <=395	
6		Area >=116 && area <=160	
7		Area >=250 && area <=295	
8		Area >=320 && <=1185	
9		Area >=265 && area <=2147	

4. CONCLUSION

In this paper, a technique for extracting diacritics from Arabic handwritten image was presented. Moreover, the proposed technique was region-based segmentation which is simplest technique for extracting diacritics. This technique is tested in Arabic handwritten document that contain many diacritic marks. Furthermore, the proposed technique consists of five steps compined together to deliver the final result. In conclusion, this technique can be improved for futher research to become more accurate. Future work is to improve the proposed technique to be able to estimate and find the best area measurement that can extract diacritics fully and accurately.

ACKNOWLEDGEMENTS

The authors thank the Ministry of Education for funding this study through the following grants: FRGS/1/2017/ICT02/FTMK-CACT/F00345. Gratitude is also due to Universiti Teknikal Malaysia Melaka and Faculty of Information Technology and Communication for providing excellent research facilities.

REFERENCES

[1] M. Jarrar, F. Zaraket, R. Asia, and H. Amayreh, "Diacritic-Based Matching of Arabic Words," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 2, pp. 1-21, 2018.
[2] A. Shatnawi and K. Omar, "Methods of Arabic Language Baseline Detection-The State of Art," *J. Comput. Sci.*, vol. 8, no. 10, pp. 137-143, 2008.
[3] M. Lutf, X. You, and H. Li, "*Offline Arabic handwriting identification using language diacritics*," Proc.-Int. Conf. Pattern Recognit., pp. 1912–1915, 2010.
[4] I. Ullah, M. S. Azmi, M. I. Desa, and Y. M. Alomari, "Segmentation of touching Arabic characters in Handwritten documents by overlapping set theory and contour tracing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 155-160, 2019.
[5] H. Bouamor *et al.*, "A Pilot Study on Arabic Multi-Genre Corpus Diacritization Annotation," *Pro-ceedings Second. Arab. Nat. Lang. Process.*, pp. 80-88, 2015.
[6] M. Farchi, K. Tahiry, S. Mounir, B. Mounir, and A. Mouhsen, "Energy distribution in formant bands for arabic vowels," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 2, pp. 1163-1167, 2019.
[7] S. Harrat, M. Abbas, K. Meftouh, and K. Smaili, "*Diacritics restoration for Arabic dialect texts*," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, no. August, pp. 1429-1433, 2013.
[8] O. Shaaban, "Automatic Diacritics Restoration for Arabic Text," *Emnlp*, vol. 12, no. September, pp. 2286-2291, 2013.
[9] M. Lutf, X. You, Y. M. Cheung, and C. L. Philip Chen, "Arabic font recognition based on diacritics features," *Pattern Recognit.*, vol. 47, no. 2, pp. 672-684, 2014.
[10] A. R. Radzid, M. S. Azmi, I. E. A. Jalil, N. A. Arbain, A. K. Draman Muda, and A. Tahir, "Text line segmentation for mushaf Al-Quran using hybrid projection based neighbouring properties," *J. Telecommun. Electron. Comput. Eng.*, vol. 10, no. 2-7, pp. 53-57, 2018.
[11] امیدعلی. ق and آذری. ا, "No Titleبیب ضریب تأثیر", vol. 8, no. 33, p. 44, 2014.
[12] A. Gutub, Y. Elarian, S. Awaida, and A. Alvi, "*Arabic Text Steganography Using Multiple Diacritics*," WoSPA 2008-5th IEEE Int. Work. Signal Process. its Appl., pp. 18-20, 2008.
[13] S. S. Bafjaish, A. Ramzani, M. Nasser, M. Sanusi, and H. Mahdin, "Skew Detection and Correction of Mushaf Al-Quran Script using Hough Transform," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 8, pp. 402-409, 2018.
[14] A. Odeh and K. Elleithy, "*Steganography in Arabic Text Using Full Diacritics Text*," 25th Int. Conf. Comput. Their Appl. Ind. Eng., no. May, 2012.
[15] H. K. Tayyeh, M. S. Mahdi, and A. S. A. AL-Jumaili, "Novel steganography scheme using Arabic text features in Holy Quran," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 3, pp. 1910-1918, 2019.
[16] L. B. Melhem, M. S. Azmi, A. K. Muda, N. J. Bani-Melhim, and M. Alweshah, "Text line segmentation of al-quran pages using binary representation," *Adv. Sci. Lett.*, vol. 23, no. 11, pp. 11498-11502, 2017.
[17] M. Hssini and A. Lazrek, "Design of Arabic Diacritical Marks," *IJCSI Int. J. Comput. Sci.*, vol. 8, no. 3, pp. 1-10, 2011.
[18] Y. Elfakir, G. Khaissidi, M. Mrabti, and D. Chenouni, "The Impact of the Image Processing in the Indexation System," vol. 9, no. 5, pp. 4311-4320, 2019.
[19] K. F. Q. P. Complex, "Digital Copy of Mus'haf Al-Madinah," 1985. [Online]. Available: https://dm.qurancomplex.gov.sa/hafsdownload.
[20] D. A. Mohammed, A. A. H. Mezher, and H. S. Hadi, "Off-line handwritten character recognition using an integrated DBSCAN-ANN scheme," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1443-1451, 2019.
[21] A. Souhar, Y. Boulid, E. Ameur, and M. Ouagague, "Segmentation of Arabic Handwritten Documents into Text Lines using Watershed Transform," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 4, no. 6, p. 96, 2017.
[22] S. S. Bafjaish, M. S. Azmi, M. N. Al-mhiqani, and A. A. Sheikh, "Skew correction for mushaf Al-Quran : a review Skew correction for mushaf Al-Quran : a review," no. September 2019, 2020.
[23] M. Kaur and P. Goyal, "A Review on Region Based Segmentation," *Int. J. Sci. Res.*, vol. 4, no. 4, pp. 2319-7064, 2013.

[24] MathWorks, "Measure properties of image regions." [Online]. Available: https://www.mathworks.com/help/images/ref/regionprops.html.
[25] K. V. Sánchez, "Functional-Communicative Grammar (Spanish-German) for Translators and/or Interpreters: A Project," *Babel*, vol. 47, no. 2, pp. 109–120, 2002.

BIOGRAPHIES OF AUTHORS

Ahmed Abdalla Sheikh received his BSc in Computer Science (Database Mangement System) in 2017 and MSc in Computer Science (Software Engineering and Intelligent) from the Universiti Teknikal Malaysia Melaka (UTeM) in 2019. His research intrests include image processing, image segmentation, computer vision systems.

Mohd Sanusi Azmi is an Associate Professor Department of Software Enginering, Universiti Teknikal Malaysia Melaka (UTeM). He received BSc. Msc and Ph.D from Universiti Kebangsaan Malaysia (UKM) in 2000, 2003 and 2013. He is the Malaysian pioneer researcher in identification and verification of digital images of Al-Quran Mushaf. He is also involved in Digital Jawi Paleography. He actively contributes in the feature extraction domain. He has proposed a novel technique based on geometry feature used in Digit and Arabic based handwritten documents.

Maslita Abd Aziz is a senior lecturer at Faculty of Information and Communication Technology, UTeM. She finished her first degree at Universiti Utara Malaysia (UUM) in 1995 with BSc in Information Technology (with Hons.) and later her MSc from Rochester Institute of Technology, New York, USA in 1998 with MSc in Information Technology (with Hons) specializing in Software Development and Management. Her research interests are in information retrieval, specifically on code retrieval of how to assist programmers during system development or learning the language

Mohammed Nasser Al-Mhiqani received his BSc in Computer Science (Computer Networking) in 2014, and MSc in Computer Science (Internetworking Technology) from the Universiti Teknikal Malaysia Melaka (UTeM) in 2015. Currently, he is a PhD student at the Universiti Teknikal Malaysia Melaka (UTeM). His research interests include cyber security, cyber-physical system security, insider threats, machine learning, and image Processing.

Salem Saleh Bafjaish received his Bachelor Degree from Staffordshire University, Malaysia in 2014 in computing (Software Engineering) and his Master Degree from UTeM university department of Information and Communications Technology (Software Engineering and Intelligent) in 2019. His current research interests are document analysis, image processing, computer vision, machine learning.