IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Optical Character Recognition for Quranic Image Similarity Matching

Faiz Alotaibi[1], Muhamad Taufik Abdullah[1], Rusli Bin Hj Abdullah[1], Rahmita Wirza Binti O. K. Rahmat[1], Ibrahim Abaker Targio Hashem[2], Arun Kumar Sangaiah[3]

[1]Faculty of Computer Science & Information Technology, University Putra Malaysia, 26600 Pekan, Pahang, Malaysia.

[2] Department of Computing Technology, Asia Pacific University of Technology and Innovation Technology, Taman Teknologi Malaysia, 57000, Malaysia.

[3]VIT University, Vellore-632014, India.

Corresponding author: Faiz Alotaibi (e-mail: faiz.eid@hotmail.com, targio_123@yahoo.com).

ABSTRACT The detection and recognition and then conversion of the characters in an image into a text are called optical character recognition (OCR). A distinctive type of OCR is used to process Arabic characters, namely, Arabic OCR. OCR is increasingly used in many applications where this process is preferred to automatically perform a process without human association. The Quranic text contains two elements, namely, diacritics and characters. However, processing these elements may cause malfunction to the OCR system and reduce its level of accuracy. In this paper, a new method is proposed to check the similarity and originality of Quranic content. This method is based on a combination of Quranic diacritic and character recognition techniques. Diacritic detections are performed using a region-based algorithm. An optimization technique is applied to increase the recognition ratio. Moreover, character recognition is performed based on the projection method. An optimization technique is applied to increase the recognition ratio. The result of the proposed method is compared with the standard Mushaf al Madinah benchmark to find similarities that match with texts of the Holy Quran. The obtained accuracy was superior to the other tested K-nearest neighbor (knn) algorithm and published results in the literature. The accuracies were 96.4286% and 92.3077% better in the improved knn algorithm for diacritics and characters, respectively, than in the knn algorithm.

INDEX TERMS Image processing, Character Recognition, Quranic diacritics, knn, optimization.

I. INTRODUCTION

character recognition (OCR) is the process of converting an image representation of a document into an editable format [1]. This application enables users to search for documents stored in the format of images by converting them into text, which can be easily performed and processed by computers. Each OCR system contains a few processing stages, in which a particular task can be accomplished, and the output of each stage is considered the input for the next stage. A typical OCR system consists of a few main stages, including preprocessing, segmentation, feature extraction, and classification. However, the ultimate goal of developing a method with the same reading capabilities as humans has remained unreached even after many years of intensive investigation and research.

Reference [2] provides a survey of Arabic OCR (AOCR) based on text recognition by focusing on the characteristics and technologies of text recognition in the Arabic language. In addition, OCR becomes an important trend in the current literature given the availability of applications that generates texts and images.

In this paper, a new, simple, typewritten, offline character recognition for the Arabic language is presented. The research on handwritten text recognition has become challenging for decades, thereby gaining increasing attention due to the increasing popularity of handheld computers, digital notebooks, and advanced cellular phones. Certain techniques, such as neural networks, hidden Markov model (HMM), and fuzzy logic, have been used to build several handwritten systems [3].

Complex character recognition algorithms have been developed after 1990. Many recognizers used sophisticated methodologies, such as neural networks, HMM, and natural language processing (NLP) techniques. However, current algorithms generate error rates that restrict full recognition for a user and comparison with the Quran. The current AOCR inaccurately recognizes diacritic and characters, and research and effort in the area of AOCR are insufficient.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2771621, IEEE Access

This paper presents a similarity check algorithm for Quranic character and diacritic recognition in the image. Moreover, a new method is proposed to check the similarity and originality of online Quranic content. The method consists of Quranic diacritic and character recognition techniques. The detection of diacritics is performed using a region-based algorithm. An optimization technique is applied to increase the recognition ratio. The character recognition method is performed based on the projection method. Then, the combined result of the two methods enables its comparison with the standard Mushaf al Madinah benchmark and finds a match on the Quran. Finally, the similarity ratio of the given image and its matching benchmark is determined.

The rest of this paper is organized as follows. Section 2 provides an overview of the OCR for Quranic image similarity matching. Section 3 presents several related works. Section 4 introduces the proposed solution based on Quranic diacritic recognition for image processing, Quranic character recognition of an image, and checking similarity matching with the standard version of the Quran. Section 5 describes the experimental setup. Section 6 analyzes the results of experimental testing. Section 7 summarizes the metrics, analysis, and comparison. Section 8 discusses the similarity check. Section 9 draws the conclusion.

II. OVERVIEW

The Arabic language is considered the most dominantly spoken language in all Arabic countries and the second language of approximately 280 million people worldwide. A recently conducted study states that the Arabic language ranks fifth as the most common language used in the world. Thus, religious beliefs and practices of Islam require that Muslims read the Holy Quran using the Arabic language, thereby requiring basic knowledge of Arabic when praying [4]. Moreover, many other languages, such as Persian, Jawi, Pashto, Urdu, and Bengali, are associated with the Arabic language; that is, certain characters are similar [5].

People can recognize characters without difficulty when reading papers or books. However, the development of an OCR system that can read and recognize Arabic Quranic characters similar to human remains unresolved [6]. Two types of OCR, namely, handwritten and typewritten, are found in the literature [7]. A typewritten OCR mainly identifies documents that are typed and scanned before being recognized. However, a handwritten OCR is used to recognize a text that is written by a human hand. The difference between the two OCR systems is that a typewritten OCR is easier than a handwritten OCR in terms of design. Moreover, the recognition rate is higher in a typewritten OCR than in a handwritten OCR.

In addition, OCRs can be further classified into two categories, namely, online and offline recognition systems [7]. The image of a typewritten or handwritten text is created via scanning using offline recognition systems. However, devices, such as a phone or a portable personal computer, are used to create an input image from the OCR system by using online recognition systems. Then, the OCR system reads the image, which is analyzed for recognition. Many other studies have been conducted in various languages, such as Japanese, Chinese, and Latin characters, which are mostly based on individual isolation of characters through OCR algorithms. However, this method may be inappropriate for languages with cursive scripts, such as Arabic. Consequently, a few studies on AOCR and its character recognition comparison with languages, such as Latin and Chinese, have been conducted [8]. Image processing has played an important role in human life. Machine learning has developed as a powerful tool for information processing, decision-making, and knowledge management. Many applications, such as digitizing libraries, postal reading, envelopes, and retrieving texts, have benefited from using OCR systems [9].

Existing algorithms and techniques have error rates that restrict a user from fully recognizing and comparing the Quran. Moreover, the diacritics and characters are recognized inaccurately. Finally, research and effort in the area of AOCR are rare. Therefore, this paper presents a similarity check algorithm for Quranic character and diacritic recognition in the image.

III. RELATED WORKS

This section provides existing techniques for Quranic character and diacritic recognition in the image. Several approaches have been introduced for Quranic character and diacritic recognition; these approaches include the segmentation of Arabic cursive script using HMM proposed by [10]. This technique operates through a morphological approach. The proposed segmentation algorithm provides satisfactory results when experimented with MM tools to test the accuracy of an algorithm for correcting and constructing slant and connected components, respectively. The image detection rate of the algorithm is approximately 81.88%. However, an improvement is required. Reference [11] introduced a segmentation-free approach to text recognition with application to Arabic text using the matching technique. This technique is based on the morphological approach to segment lines into characters and aims to implement a system that provides satisfactory results in the noise-free and synthetically degraded text with minimal efforts. However, its recognition rate is approximately 95.6% for noise-free words and synthetically degraded words and 73% for scanned words. An offline handwritten Arabic character segmentation algorithm (ACSA) based on the morphological analysis of word contours was implemented by [12]. The algorithm applies contour representation to recognize segmentation points. This approach aims to ensure that each segment contains exactly one character. However, morphological analysis or segment extraction procedure score in the approach is high. Reference [13] introduced an AOCR system using recognition-based segmentation technique. This system is a real-time process for fragmenting Arabic words using their

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2771621, IEEE Access

structural properties, connectivity points, and A copy detection pattern CDPs. The recognition rate detected with 20 characters is 90%. A new approach for Latin/Arabic character segmentation was proposed by [14]. The proposed method used the segmentation technique to solve the difficulty of overlapping characters by applying a contour-following algorithm, in which the detected contours will be labeled. In the first proposed method, the junction segments between characters are detected. The second segmentation technique uses an upper contour for each word to determine the characteristic strokes, which characterize consistent information for the segmentation of each character.

Similarly, Reference [15] uses a projection method to implement the recognition of Arabic characters. The approach uses the measurements and relationships of moments as features and a Bayesian classifier for classification. The classification rates when linear discriminant analysis and quadratic discriminant analysis are used are 85.5% and 99.5%, correspondingly. Reference [16] proposed a novel algorithm for the AOCR using neural networks to classify features. This algorithm creates tokens that characterize the characters. The proposed method mainly depends on extracting a set of features for each character and then provides all the extracted information to the recognition and assembly phases. However, the average recognition rate is only 87%. Reference [7] uses an algorithm based on neural network for recognizing Arabic text. The classifier algorithm specifies the corresponding character and the similar contours of many Arabic characters. However, only 73% of the contours are recognized. Furthermore, an Arabic handwriting recognition system that uses baseline-dependent features and HMM is proposed. The technique adopts a detection technique for extracting a set of language-independent features. These features depend on the baseline; that is, any minor error in detecting the baseline location will be problematic for the feature extraction stage. However, the experimental result shows a significant improvement in the recognition rate using baseline-dependent features; the recognition rate is approximately 86.51%.

IV. PROPOSED METHOD

This section provides the proposed technique for achieving an efficient OCR for Quranic image similarity matching. The proposed method is divided into three (3) phases of processing once the image is loaded into the system. These phases are Quranic diacritic image recognition, Quranic character image recognition, and checking similarity matching. Figure 1 illustrates the processes involved in the OCR for Quranic image similarity matching.

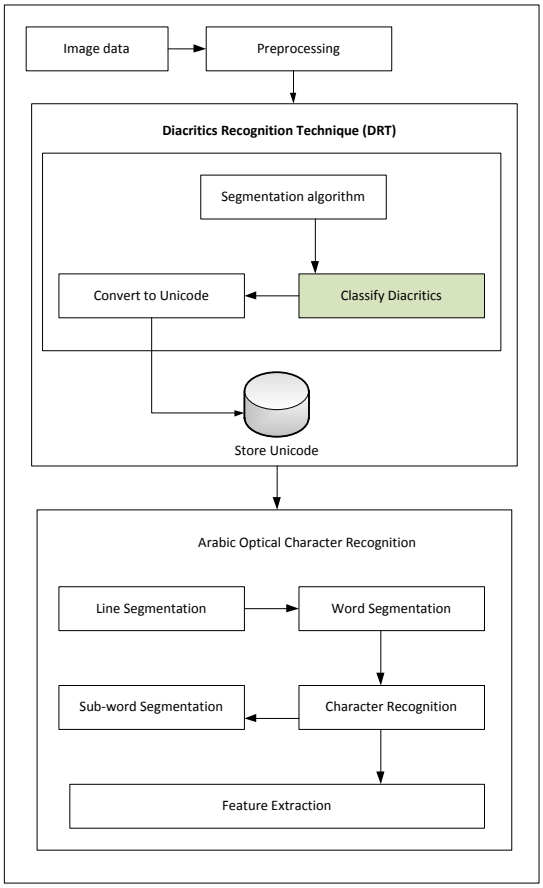

FIGURE 1. Proposed method of OCR for Quranic image similarity matching

The proposed method functions as follows. A Quranic image related to the research is collected from Internet sources. First, the image is loaded for preprocessing. Then, the diacritics of the ayah are detected using a segmentation algorithm to classify and convert the diacritics into Unicode. Second, the characters of the ayah are recognized using line and word segmentation. Finally, the features of the character are extracted based on the segmentation. The results are compared with the original standard Mushaf al Madinah; the similar part of the standard version is highlighted, and the similarity ratio is calculated and determined. The ultimate goal of the proposed Quranic image similarity matching is to optimize the recognition performance of Quranic diacritics and characters. Thus, an analysis of Quranic diacritics and characters on the basis of accuracy and correctness is conducted.

A. Quranic Diacritic Recognition

The use of baseline method in preprocessing the characters in an AOCR system can be separated from diacritics. The process of Arabic characters is considered complicated because the language comprises many cursive alphabets. The diacritics are separated from Arabic characters by removing the letter from the image, as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2771621, IEEE Access

demonstrated in Figure 2.

FIGURE 2. Process of removing letters from image

Diacritic recognition is conducted by undergoing a series of stages/steps, as depicted in Figures 3 and 4. The first part is training the images. This part begins by inserting all the images and detecting all the diacritic inputs. Then, two tasks, namely, feature extraction for all diacritic and manual label assignment for all diacritic, are performed. Thereafter, the datasets are trained. K-nearest neighbor (knn) algorithm is applied during the training. knn is a generalization of the classical knn classifier, and the result is saved as knn model. During the testing, the knn model that is generated during the training is loaded, and the text images are inputted. In this stage, all the diacritics are detected as an input, and a featured extraction for all diacritics is performed based on the trained model and the manual evaluation of the label. Table 1 summarizes the feature extraction algorithm using 2D maximum embedding difference.

Table 1. 2D Maximum Embedding Difference Algorithm

```
% input: all training images
% output : optimum scatter matrix
% find matrix of distance
for i=1:M
for j=i+1:M
Dist(j,i)=Dist(i,j)=norm(TrainData(i)-TrainData(j));
end
End
% find all nearest neighbors
for i=1:size(Dist,1)
[Vals,Indx]= sort(Dist(i,:)); KNNc(i,:)=Indx;
End
NplusKc=KNNc(:,[2:Kc+1]);
End
```

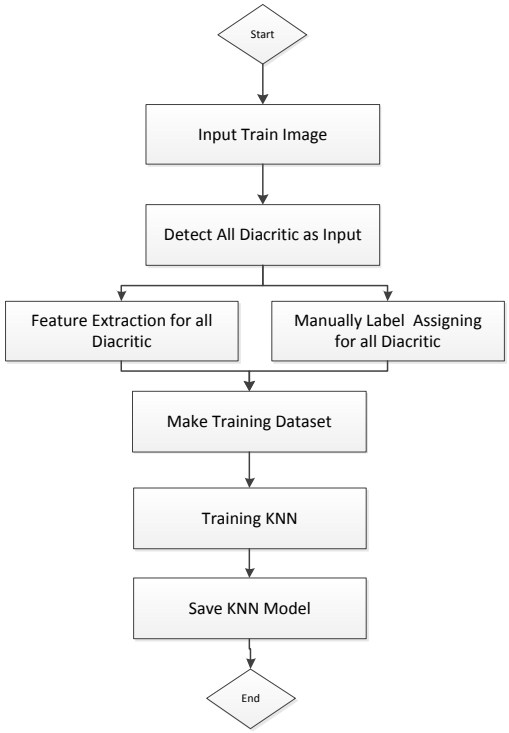

FIGURE 3. Diacritic recognition flow (train image)

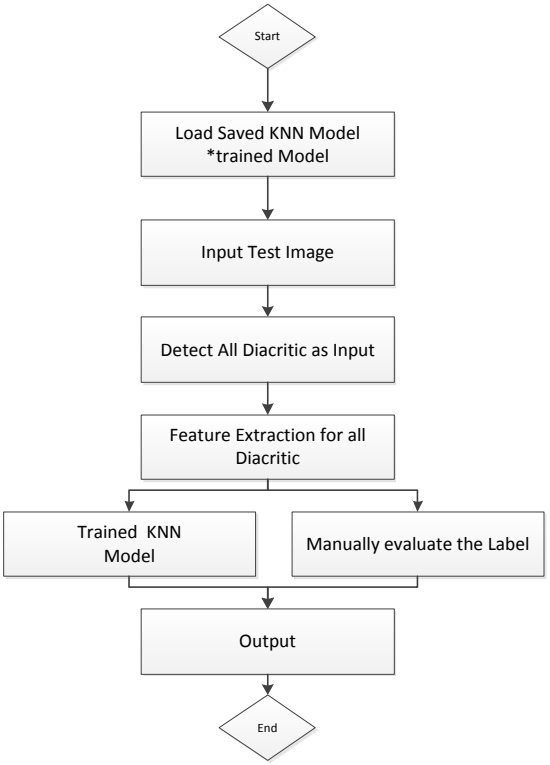

FIGURE 4. Diacritic recognition flow (test)

1. Identifying the Required Features

The possible errors that might be encountered in this phase require the identification and use of suitable features. For example, the words connected to the baseline are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2771621, IEEE Access

identified and removed to focus on a diacritic. The baselines retain the diacritics, tajweed, and other recitation symbols. In certain cases, the whole letter environment might be unrecognized as one region given the low quality of the image, as presented in Figure 5. This issue requires the use of additional features to recognize and remove non-diacritic items.

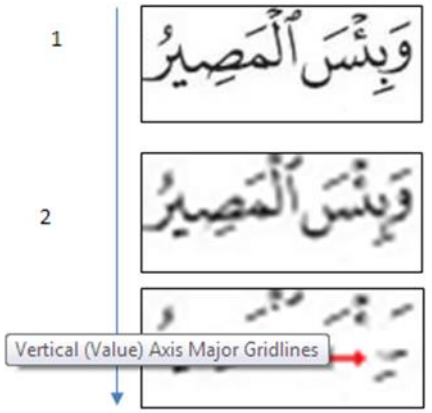

FIGURE 5. Error occurred after removing extra letters

2. *Distinguish the Type of Each Diacritic*

Each type of Arabic diacritics has a distinctive shape, except the Fatḥah and Kasrah. The two shapes can be recognized from their locations from the baseline. If the shape is located on the upper line, then the diacritic is Fatḥah; otherwise, the diacritic is Kasrah. An algorithm is developed to calculate the area of each type of diacritic. The algorithm aims to recognize the region of each diacritic from the binary image. Thus, a fictional elliptic is sketched, and the shape of the diacritic can be recognized by the environment distance of each diacritic from the border of the ellipse.

3. *Quranic Character Recognition*

The use of the baseline method in preprocessing the characters in the AOCR system can be separated from diacritics. The process of the Arabic characters is considered complicated because the language consists of many cursive alphabets. This separation can ease the examination of the Arabic language, as visualized in Figure 6.

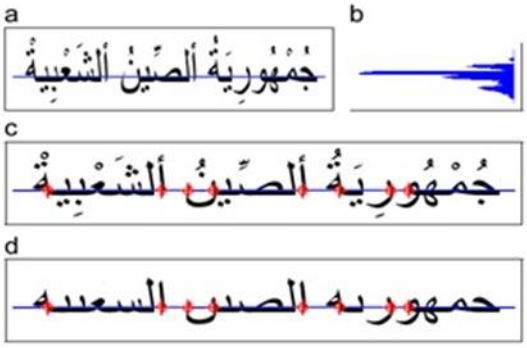

FIGURE 6. Process of removing diacritics from image

Figure 7 displays the preprocessing stage of the AOCR. The improvement in the preprocessing stage in the AOCR has resulted in an accurate feature extraction for each character. The location of the line with characters is precise when the line segmentation process is enhanced. The word and sub-word segmentation algorithm should be required to increase the efficiency of detecting Arabic characters.

FIGURE 7. AOCR Preprocessing Stage

4. *Identifying the Required Features*

The use of a baseline to segment text into words or characters is beneficial because the Arabic language is a cursive. Moreover, this method is advantageous for extracting dependent features. Arabic characters are categorized into two groups, namely, ascenders and descenders, through the baseline method. These groups may be constructed from stroke or small element or complete character. This Arabic language features can cause complexity, such as the possibilities of combinations, for recognizing many Arabic characters.

5. *Checking Similarity Matching with Standard Version of the Quran*

The removal of certain noise and enhanced accuracy of filters are required after converting a given image into text through the proposed AOCR system. In this phase, the proposed system output will be compared with a standard

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2771621, IEEE Access

version of the Quran to find a matching phrase with the genuine Quran. Several errors might occur while performing the matching. An artificial intelligence technique for NLP is used to increase the efficiency of the proposed method.

V. EXPERIMENTAL SETUP

This section describes the experiments conducted to compare and evaluate the recognition of Arabic characters and diacritics. A 2D maximum embedding difference feature extraction is used for the performance evaluation of Arabic characters. However, the recognition of diacritics is based on a known algorithm. The experiment is conducted using MATLAB software on a 32-bit Windows 7 Professional machine. The processor used is Intel Pentium, 2.20 GHz, and 4.00 GB. A total of 10 different experiments are conducted to detect a Quranic image from the online source.

VI. RESULTS OF EXPERIMENTAL TESTING

The results are presented in percentages based on the accuracy rate for Quranic diacritics and characters. The result in Figure 8 indicates that the improved knn algorithm has an average of 96.4286% accuracy performance, which is better than the normal knn for diacritic recognition. The comparison of the discussed algorithms revealed that the improved knn has significantly improved the accuracy.

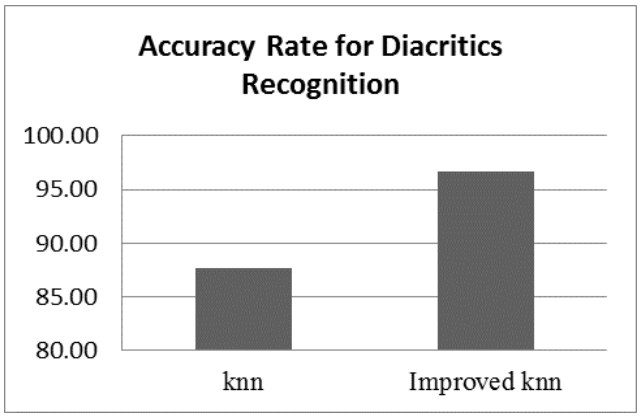

FIGURE 8. Accuracy of the diacritics

The results in Figure 9 denote the percentage of the accuracy rate based on character recognition. The results confirm that the improved knn algorithm has achieved an accuracy of 92.3077% compared with the knn, which has a 78.8022% accuracy. The main purpose of achieving high accuracy is to increase the efficiency of detecting Arabic characters.

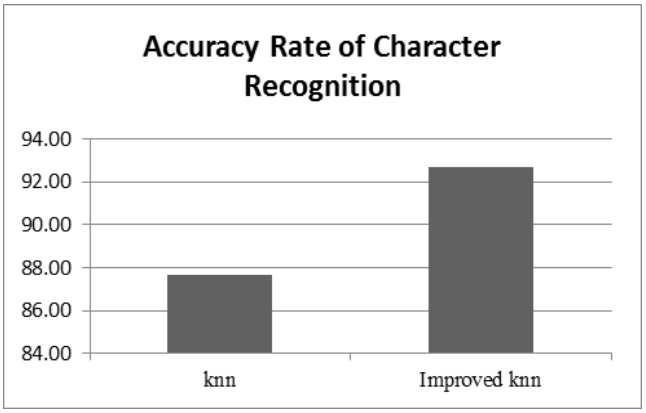

FIGURE. Accuracy of the character recognition

VII. METRICS, ANALYSIS, AND COMPARISON

This section presents the metrics, such as precision, recall, and f-measure, for measuring and justifying the validity of the results in our experiments using the common performance measures from the information retrieval (IR) domain. These measurements are used to offer a significant insight into the performance characteristics of our classification as discussed below:

Precision is the ratio of correctly detected class to the number of overall detected classes.

$$\text{Precision} = \frac{t_p}{t_p + f_p},$$

where t_p is the number of true positive, and f_p is the number of false positive sentiments.

$$recall = \frac{t_p}{t_c},$$

where t_c is the overall number of classes detected in the sentiment analysis. F-measure can be presented in terms of precision and recall as follows.

$$\text{F} - \text{measure} = \frac{2\,x\,precision\,x\,recall}{precision + recall}$$

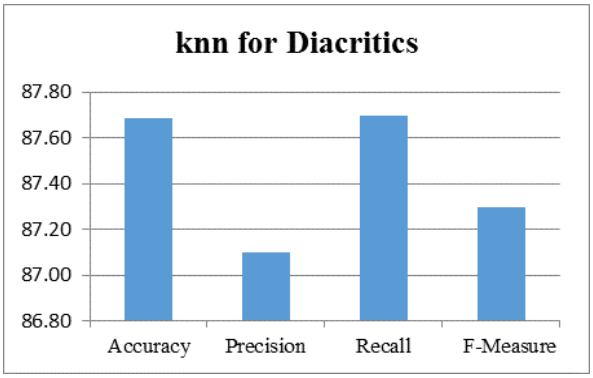

FIGURE 10. Precision, recall, f-measure, and accuracy of experiments of the diacritic recognition using the knn algorithm

Figure 10 illustrates that the number of accuracies, recall, f-measure, and precision was applied to the experiment to justify the validity of their results. The result indicates that recall has the highest value compared with precision and f-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2771621, IEEE Access

measure, which has low values when the knn is used. Moreover, the high accuracy in detecting the alterations of the sample and the standard version are depicted in Figure 10.

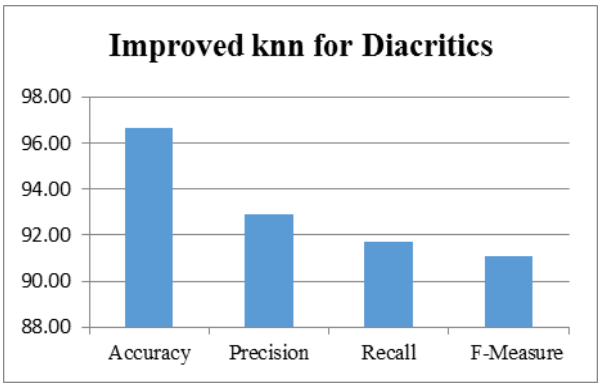

FIGURE 11. Precision, recall, f-measure, and accuracy of the experiments of the diacritic recognition using the improved knn algorithm

However, Figure 11 displays that precision scores are high because of numerous correctly identified Quranic images compared with the recall and f-measure, in which the diacritic method is applied. The measurements for the set of properties are specified based on the properties of each connected component (object) in the binary image. The comparison of the knn and improved knn algorithms based on precision, recall, f-measure, and accuracy are presented in Figure 12. The improved knn has high accuracy and correctness detection. Moreover, the metrics are transformed into precision, recall, and F-measure based on our experimental results.

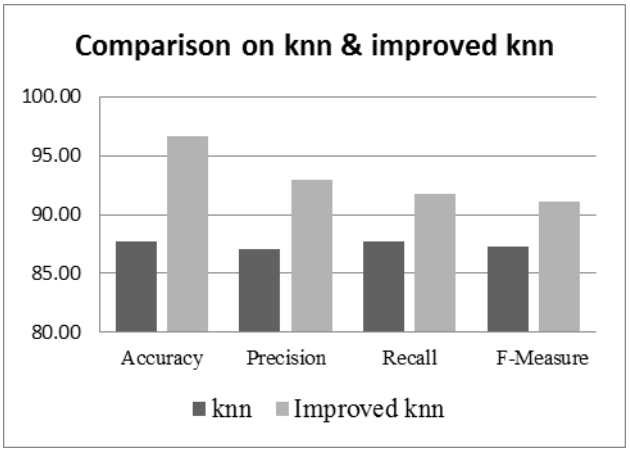

FIGURE 12. Precision, recall, f-measure, and accuracy of experiments of the diacritic recognition through the knn algorithm

VIII. DISCUSSION ON SIMILARITY CHECK

The results of diacritic recognition technique (DRT) and AOCR are presented in this section. In this research, the DRT preprocessing stage should be enhanced to improve the accuracy of detecting a diacritic. Furthermore, the final result of diacritic recognition is significantly enhanced by improving the efficiency of the segmentation algorithm. The DRT uses "OCR to detect diacritic marks in detected text. The method includes; receiving, by a computer system, an electronic image containing text including a plurality of diacritics; analyzing, by the computer system, the electronic image to generate a plurality of bounding blocks associated with text within the electronic image, wherein the plurality of bounding blocks include at least a base text bounding box and a diacritic bounding box; determining a base box distance between the diacritic bounding box and a nearest base text bounding box; analyzing the plurality of bounding blocks to determine a plurality of text lines; determining a proximity value for the diacritic bounding box to a nearest text line of the plurality of text lines; associating, by the computer system, the diacritic bounding box with a corresponding text line based on the proximity value and the base box distance, whereby the diacritic bounding box association is thus made responsive to a determination of multiple distance values, namely the proximity value and the base box distance; and processing the plurality of bounding blocks to produce electronic text from the electronic image", as displayed in Figure 13.

FIGURE 13. Diacritic recognition using OCR

The results of the AOCR in this research are achieved in phases. The phases are as follows: first, the preprocessing stage involves image reading, denoising (if necessary), resizing (if necessary), conversion to gray, conversion to binary using threshold, and complementation. Second, the line segmentation stage detects the total number of lines and images. Third, the "word segmentation" stage involves segmenting all words of one line and repeating the same process for all lines in the image, as depicted in Figure 14.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2771621, IEEE Access

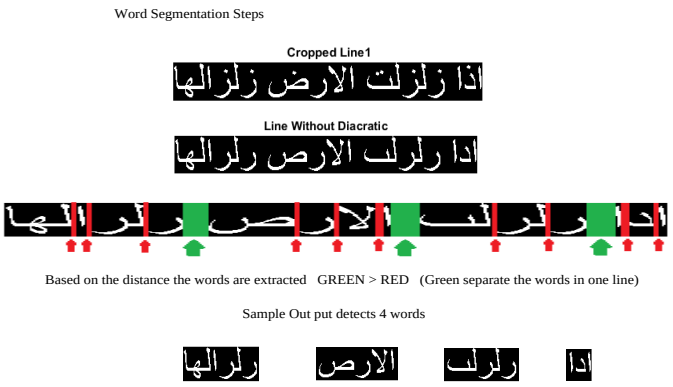

FIGURE 14. Arabic word segmentation

Figure 14 illustrates Arabic word segmentation. Each segmented word is further segmented after the segmentation stage. This phase is called "sub-word segmentation." The following tasks are performed in the sub-word segmentation stage: all sub-words of one word are segmented. This process is repeated for all words in one line. A threshold is not required for the segmentation, which is only based on the distance between each part with and without diacritic segmented sub-words, as presented in Figure 14.

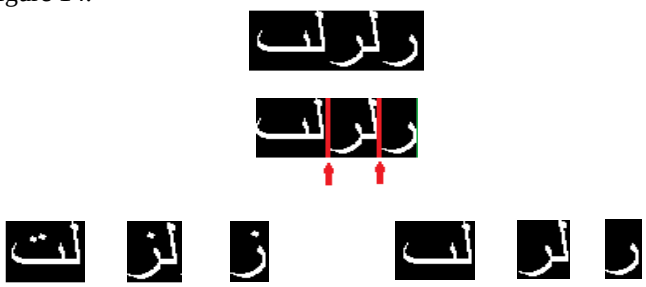

FIGURE 15. Arabic sub-word segmentation

Figure 15 displays the Arabic sub-word segmentation. An Arabic word is either a capital letter or a small letter that can be sub-segmented. Capital and small letters are detected and recognized after the sub-word segmentation stage. The remaining character of each sub-word and the lone characters are capital. This process is repeated for all sub-words to separate the capital characters from the rest of the sub-words. The character detection and recognition process (capital and small letters) are demonstrated in Figure 16.

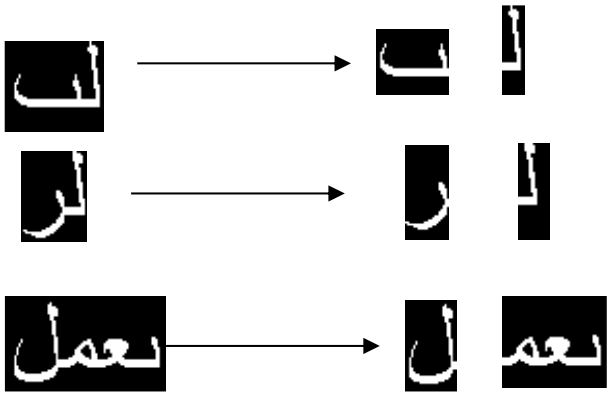

FIGURE 16. Arabic capital and small letter detection and recognition

The detection and recognition of capital and small Arabic letters are presented in Figure 14. If the character is a capital letter, then the capital recognition dataset is used for recognition. If the character is a small letter, then a "sliding window" is used to recognize the small characters individually. The method of recognition is different. In the recognition phase, the predictive model is used.

IX. CONCLUSION

Image processing is considered an emerging area of research in the field of computer science. The management of images that contain noise and distortions through traditional methods in image processing is challenging. Thus, computational intelligence approaches, such as knn, have been recently used by many researchers to address the aforementioned challenges. Computational intelligence approaches are suitable when the data to be modeled are complex for conventional statistical techniques for rapid and efficient processing. In this research, a new method was proposed to check the similarity and originality of online Quranic content. This method is a combination of Quranic diacritic and character recognition techniques. The diacritic detections are based on a region-based algorithm. An optimization technique is applied to increase the recognition ratio. The character recognition method is based on the projection method. Moreover, an optimization technique is applied to increase the recognition ratio. The combination of the result of the two methods was compared with the standard knn, and the match in the Quran was found. Then, the similarity ratio of the given image and its matching benchmark is determined. The obtained accuracy was superior to the other tested knn algorithm and published results in the literature. The accuracy was 96.4286% better in the improved knn algorithm than in the knn algorithm with 87% accuracy in detecting diacritic and 92.3077% better in the improved knn algorithm than in the knn algorithm, which exhibited an 86% accuracy in detecting the characters. This can be used to recognizing and parsing a medical image into multiple objects and structures.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2771621, IEEE Access

Many challenges have not been addressed, although the applications of knn have resulted in the improvements in image processing. Thus, future opportunities in the area of computational intelligence algorithm applications, such as deep neural network and support vector machine, and deep learning can be applied. Moreover, artificial neural network learning rules can also be adopted through evolutionary algorithms.

REFERENCES

[1] A. M. Al-Shatnawi, "A new method in image steganography with improved image quality," *Applied Mathematical Sciences*, vol. 6, no. 79, pp. 3907-3915, 2012.

[2] H. A. Al-Muhtaseb, S. A. Mahmoud, and R. S. Qahwaji, "Recognition of off-line printed Arabic text using Hidden Markov Models," *Signal processing*, vol. 88, no. 12, pp. 2902-2912, 2008.

[3] V. Vuori, M. Aksela, J. Laaksonen, E. Oja, and J. Kangas, "Adaptive character recognizer for a hand-held device: Implementation and evaluation setup," in *Proc. of the 7th IWFHR*, 2000, pp. 13-22.

[4] Z. Fan, D. Bi, L. He, M. Shiping, S. Gao, and C. Li, "Low-level structure feature extraction for image processing via stacked sparse denoising autoencoder," *Neurocomputing*, vol. 243, pp. 12-20, 2017.

[5] M. A. Abuzaraida, A. M. Zeki, and A. M. Zeki, "Segmentation techniques for online Arabic handwriting recognition: a survey," in *Information and Communication Technology for the Muslim World (ICT4M), 2010 International Conference on*, 2010, pp. D37-D40: IEEE.

[6] C. Shanthi and N. Pappa, "An artificial intelligence based improved classification of two-phase flow patterns with feature extracted from acquired images," *ISA transactions*, vol. 68, pp. 425-432, 2017.

[7] M. Khemakhem and A. Belghith, "A multipurpose multi-agent system based on a loosely coupled architecture to speedup the DTW algorithm for Arabic printed cursive OCR," in *Computer Systems and Applications, 2005. The 3rd ACS/IEEE International Conference on*, 2005, p. 121: IEEE.

[8] A. Mesleh *et al.*, "An optical character recognition," *Contemporary Engineering Sciences*, vol. 5, no. 11, pp. 521-529, 2012.

[9] S. Hakak, A. Kamsin, O. Tayan, M. Y. I. Idris, A. Gani, and S. Zerdoumi, "Preserving Content Integrity of Digital Holy Quran: Survey and Open Challenges," *IEEE Access*, 2017.

[10] D. Motawa, A. Amin, and R. Sabourin, "Segmentation of Arabic cursive script," in *Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on*, 1997, vol. 2, pp. 625-628: IEEE.

[11] B. Al-Badr and R. M. Haralick, "Segmentation-free word recognition with application to Arabic," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1995, vol. 1, pp. 355-359: IEEE.

[12] T. Sari, L. Souici, and M. Sellami, "Off-line handwritten Arabic character segmentation algorithm: ACSA," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, 2002, pp. 452-457: IEEE.

[13] A. Cheung, M. Bennamoun, and N. W. Bergmann, "Implementation of a statistical based Arabic character recognition system," in *TENCON'97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications., Proceedings of IEEE*, 1997, vol. 2, pp. 531-534: IEEE.

[14] K. Romeo-Pakker, H. Miled, and Y. Lecourtier, "A new approach for Latin/Arabic character segmentation," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1995, vol. 2, pp. 874-877: IEEE.

[15] H. Al-Yousefi and S. Udpa, "Recognition of Arabic characters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, pp. 853-857, 1992.

[16] M. Bokser, "Omnidocument technologies," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1066-1078, 1992.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2771621, IEEE Access

Faiz Alotaibi is currently a Ph.D candidate at the department of Computer Science & Information Technology, Unversity Putra Malaysia. His area of research includes, Information retrieval, Image processing and big data.

Muhamad Taufik Abdullah is an associate professor from the Department of Multimedia, Faculty of Computer Science and Information Technology, University Putra Malaysia. He obtained doctor of philosophy from Universiti Putra Malaysia. He joined Universiti Putra Malaysia since 1990 as tutor. His research interest includes Information Retrieval, Cross Language Information Retrieval and Multimedia Computing

Professor Dr. Rusli Abdullah is currently a Professor and Leader of Applied Informatics Research Group (AIRG) in the Department of Software Engineering and Information Systems at Universiti Putra Malaysia (UPM). He holds a PhD from UniversitiTeknologi Malaysia (2005), Master Science in Computer Science (1996) and Bachelor in Computer Science (1988) from UPM. He is also an active member of Association for Information Systems (AIS). For the past twelve years, his major research interests lie in knowledge management and information systems. He has authored and co-authored over 80 journals and 63 prestigious conferences.

RahmitaWirza O.K Rahmat, Ph D. obtained her B.Sc. and M.Sc. degrees in Science Mathematics from University Science Malaysia in 1989 and 1994 respectively. She received her PhD in Computer Assisted Engineering from University of Leeds, U.K. She is currently a Professor at Faculty of Computer Science and Information Technology, UPM.

Ibrahim Abaker Targio Hashem is a lecturer a Ph.D at the Department of Computer Systems and Technology, APU University, Kuala Lumpur, Malaysia; he received his PhD from University of Malaya , M.S. degree in computing in 2012, Malaysia, and the B.E. degree in computer science in 2007, Sudan. Hashem obtained professional certificates from CISCO (CCNP, CCNA, and CCNA Security) and APMG Group (PRINCE2 Foundation, ITIL v3 Foundation, and OBASHI Foundation). His main research interests include big data, cloud computing, distributed computing, and network.

Arun Kumar Sangaiah has received his Master of Engineering (ME) degree in Computer Science and Engineering from the Government College of Engineering, Tirunelveli, Anna University, India. He had received his Doctor of Philosophy (PhD) degree in Computer Science and Engineering from the VIT University, Vellore, India. He is presently working as an Associate Professor in School of Computer Science and Engineering, VIT University, India. His area of interest includes software engineering, computational intelligence, wireless networks, bio-informatics, and embedded systems. He has authored more than 100 publications in different journals and conference of national and international repute. His current research work includes global software development, wireless ad hoc and sensor networks, machine learning, cognitive networks and advances in mobile computing and communications. He is an active member in Compute Society of India. Moreover, he has carried out number of funded research projects for Indian government agencies. Also, he was registered a one Indian patent in the area of Computational Intelligence. Besides, Prof. Arun Kumar Sangaiah is responsible for Editorial Board Member/Associate Editor of various international journals like International Journal of Intelligent Information Technologies (IGI), International Journal of Cloud Applications and Computing (IGI), International Journal of High Performance System (Inderscience), International Journal of Image Mining (Inderscience), International Journal of Intelligent Engineering and Systems, International Journal of Computational Systems Engineering (Inderscience) and Institute of Integrative Omics and Applied Biotechnology (IIOAB), etc. In addition, he has edited number of guest editorial special issues for various journals like Applied Soft Computing, Computers and Electrical Engineering (SCI) Future Generation Computer Systems (SCI), Neural Network World (SCI), Intelligent Automation & Soft Computing (SCI), Scientific World Journal (SCI) etc. Also, he has organized a number of special issues for Elsevier, Inderscience, Springer, Hindawi, and IGI publishers etc. Also he has acted as a book volume editor of various publishers for Taylor and Francis, Springer, IGI, etc. Furthermore, Prof. Sangaiah made outstanding efforts and contributions on the technical programme committee member of various reputed international/national conferences.