International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S3, September 2019

A Method for Arabic Handwritten Diacritics Characters

Faiz Alotaibi, Muhamad Taufik Abdullah, Rusli Abdullah, Rahmita Wirza, Masrah Azrifah Azmi Murad

Abstract*: *An Optical Character Recognition (OCR) is the process of converting an image representation of a document into an editable format. In addition, people have the ability to recognize characters without difficulty as reading papers or books. However, developing an OCR system that has the ability to read and recognized Arabic diacritics characters as human still, remain a problem. More, specifically, poor recognition rate in most of optical diacritics characters recognition is mainly attributed to failing in segmenting a handwritten text correctly. To overcome this problem, we perform develop a method based on seven operations; it starts with searching the text-line height followed by reading words from the line. Then identify the diacritics regions. The segmentation is also applied during this operation by converting the text-line into a grayscale and binary image. Moreover, we introduced a new model based on k-nearest neighbors (KNN) algorithm to identify diacritics and characters segmentation. KNN is trained to directly predict the diacritic from the text-line. Finally, we offer an evaluation discussion on optical diacritics characters recognition.

Index Terms*: *diacritics characters, Handwritten, KNN, Image recognition.

I. INTRODUCTION

An Optical Character Recognition (OCR) is the process of converting an image representation of a document into an editable format (Al-shatnawi, 2012). These applications also enable 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 few processing stages, a particular task can be accomplished in each stage and the output of each stage is considered as the input for the next stage (Schantz, 1982). Commonly an OCR system consists of few main stages which are including preprocessing, segmentation, feature extraction and classification (Singh, Khan, Bansal, & Bansal, 2015). However, after many years of intensive investigation and research, the ultimate goal of developing a method with the same reading capabilities as humans still remains unachieved especially a language like Arabic (Smith, 2007). Based on the study conducted recently, Arabic language is ranking number fifth as the most common language used in the world. There are three main types of the Arabic language. These types are Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialect Arabic (DA). (AlOtaibi & Khan, 2017), (Ibrahim, Abdou, & Gheith, 2015) (i) the Classical Arabic (CA) is the oldest version of Arabic, which is used in the earliest age of Arab nationalism (ii) the Modern Standard Arabic (MSA) is the formal Arabic language, which is used nowadays in education, books, newspapers, media, and even as the official language of Arabic countries (iii) dialect Arabic (DA) is a kind of colloquial language that differs from region to region in Arab countries. Moreover, there are many other languages associated with the Arabic language which has some similar characters such as Persian, Jawi, Pashto, Urdu, Bengali, etc. (Abuzaraida, Zeki, & Zeki, 2010). The typewritten is mainly to identify documents before recognizing them that are typed and scanned. However, handwritten OCR is used to recognize text that is written by human hand. The different between these two OCR systems is Typewritten OCR is easier compare to handwritten OCR in terms of design. Moreover, the recognition rate of the typewritten is higher than handwritten. People have the ability to recognize characters without difficulty as reading papers or books. However, developing an OCR system that has the ability to read and recognized Arabic Quranic characters as human still, remain unsolved. More, specifically, poor recognition rate in most of optical character recognition system OCR is mainly attributed to failing in segmenting a handwritten text correctly, regardless of how well the previous and following stages are designed (Jayech, Mahjoub, & Amara, 2016). We proposed a method to identify Quranic diacritics and characters segmentation. The method consists of Quranic diacritics techniques. The diacritics detections performed using KNN algorithm.

The rest of this paper is organized as follows. Section 1 provides an introduction. Section 2 presents related works. Section 3 provides the diacritics detections frameworks. Section 4 offers a discussion of the proposed frameworks. Section 5 draws a conclusion.

Manuscript published on 30 September 2019.
***** Correspondence Author (s)
Faiz Alotaibi, Faculty of Computer Science and Information Technology, University Putra Malaysia
Muhamad Taufik Abdullah Faculty of Computer Science and Information Technology, University Putra Malaysia.
Rusli Abdullah, Faculty of Computer Science and Information Technology, University Putra Malaysia
Rahmita Wirza, Faculty of Computer Science and Information Technology, University Putra Malaysia
Masrah Azrifah Azmi Murad, Faculty of Computer Science and Information Technology, University Putra Malaysia

II. Related Works

The authors of the accepted manuscripts will be given a copyright form and the form should accompany your final submission Extracting handwriting features can result in using approaches for segmenting individual characters and numerals by identifying the extreme points, such as the directions of left, right, top, and bottom, in both training and testing documents. Features are also extracted from images (El Kessab, Daoui, Bouikhalene, & Salouan, 2015). The extraction must be able to identify the main useful structures that occur from the relationships among successive measurement values over time.

For simple Arabic feature extraction in the recognition process, drawn in (Luqman, Mahmoud, & Awaida, 2015), for the case of Arabic font based on diacritics, steps start with diacritic segmentation. Diacritic segmentation is claimed to have zero errors, with an approximate accuracy of 98.73%, when the line split is not well separated or the directions are not portals, The methods rely on using K-Means clustering. The key benefit of the k-means process is its simplicity in terms of time and memory. Its concept is also an efficient approach because it is cited and inspired by many researchers in different fields. The output of each iteration of k-means is

necessary to pass through the validation process (Ma, Xu, Zhang, & Li, 2015).

Referring back to the terms of big data and large scale, segmentation is the main keyword in (Spera, Tegolo, & Valenti, 2015), in which useful techniques for the segmentation stage are given, as well as the capillaroscopic features for each pixel and dimensionality size to categorize digits in the “MNIST32 dataset.” Size reduction approaches are dedicated to optimizing the performance of SVM classifier and its learning on enamor datasets. Another accessible research was provided in (Khoshnevisan et al., 2015), in which customer relationship-clustering systems were used. Several approaches were proposed to associate the performances of Dimensionality and Reduction approaches in consideration of particular problems. For the genetic factor expression setting.The datasets normally contain of a minor instance of (very) high-dimensional feature vectors. Then the quantity of works challenges the assignment of feature extraction, in which novel features are recreated as a (non-invertible and perchance nonlinear) function of the main feature to improve the system performance.

In reference to (Di Martino, Hernández, Fiori, & Fernández, 2013), the classifier rules are redefined based on the traditional KNN basics; it is a hybrid between neighborhood model and decision model by including different feature sizes. This model is implemented on the “ISOLET database,” which comprises over 617 features as attributes for visualizing diverse pronunciations in English alphabets and inferring correct alphabets associated with correct features. For this case (Aksoy), a short time is needed to deal with varied features in non-stationary signals. The basic concept consists of adjusting SVMs; each percent of the time window is assigned together by including an additional term to optimize the functionality of SVM and achieve an optimum classifier. Another explanation deals with the classicalness of SVM learning, which has been tested over large and diverse datasets (up to 250,000 features and 40 different dimensions), as presented in (Kotsiantis, Zaharakis, & Pintelas, 2007), and with enamor datasets and implemented by the Fisher decision tree approach proposed in (Bennet, Ganaprakasam, & Kumar, 2015).

III. Proposed Diacritic Recognition Method

This section offers details about the proposed technique for achieving an efficient Quranic diacritics and characters segmentation. Normally in Arabic script, the characters word/subword are combined together at an imaginary line named base-line. However, in some situation the characters may join together in a separate way of the base-line to overcome this problem, we perform seven main operations; it starts with searching the text-line height followed by reading words from the line. Then identify the diacritics regions. The segmentation is also applied during this operation by converting the text-line into grayscale and binary image.

As shown in Fig. 1 the fill-in function is performed for the image data collected. We apply a diacritics recognition technique. This technique is done by searching the text-line height then we identify the diacritics regions. After diacritics regions is identified, the segmentation is performed in order to classify the diacritics and stored them in Unicode.

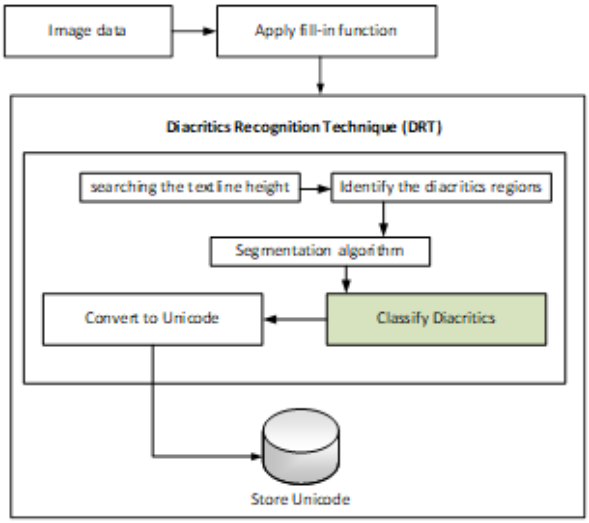

Fig. 1: Diacritic Proposed model

A. Diacritical search

Estimate the reasonable augmentation width and height using the following rule:

$$argment_{h,w}\ \text{TP} - \in \ . FP$$

Where h, w are the height and weight, TP, FP are the true positive and false positive. The number of captured diacritics results is True positive as well as the number of captured words false positive needs to be evaluated for each of the parameter vector.

IV. Segmentation of the Diacritical Search

In this study contour following-like technique is used by adopting fill-in function “FloodFill” (from VCL library) to identify all connected pixel groups (which might be words, subwords, diacritics or noise).

The fill-in function utilizes 8-connected pixels property to allocate all connected pixels of an image, in general, and we use it, in particular, to allocate connected pixels of words/subwords, diacritics or noise in a text image. The objective of the connected pixel analysis in an image is to form rectangles around distinct components in the text-line image and segment them sequentially as shown in Fig. 2.

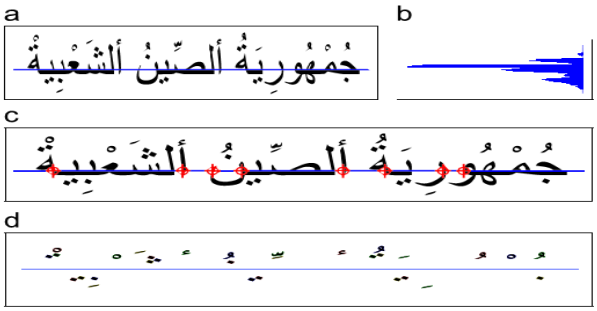

Fig. 2: text-line image segmentation

V. ADOPTED DETECTION OF DIACRITICAL

We introduce an algorithm for detecting diacritical based on search and segment method. The idea is to identify the text-line which contain the diacritical from the image. The only diacritics are extracted from the image and converted to text as described below.

1. Begin with i=0, identify h, w of the text-line
2. Search the text line image and detect the first black pixel and mark it as (X_i , Y_i)
3. Check if the black pixel is existed? If No check step 6
4. Make i=i+1, from (X_i , Y_i) apply the fill-in function, and find the coordinates of the rectangle bounding currently connected pixel (resultant shape) as a top right coordinate. If h or w is <= 3pixel, {Cut the resultant shape and go to step-2} // considered as noise Cut the resultant shape and paste it in a different array or file. Name this file Fi, calculate aspect ratio as Ari= h/w
5. Go back to (X_i , Y_i) and go to step-2
6. Find the highest value of aspect ratio (Ari) and retrieve Fi (file corresponding to that Ari)
7. Scan image Fi from top right to bottom left {Count the black pixel in the current column and stop if a white pixel is detected. If the count is >= 0.2 height, discard current column and jump to the next column. Otherwise, store the count in an array element and jump to the next column} Get the most frequent element (but not 0) in the array and make it = width width= width+0.25width
8. Initiate i=0, scan the text-line image from top right to bottom left and detect the first black pixel and mark it as (X_i , Y_i)
9. Is there a black pixel If NO, go to END

In this study we have design a flow of the diacritical detection to explain the steps involved in recognizing diacritics from images as shown in fig. 3. First we search for the lines containing the words and read the words from the line. Then, we identify the diacritical diacritics regions and convert them into a greyscale. After identifying the regions and convert them into grayscale, we then covert the results into binary images in order to segmentation.

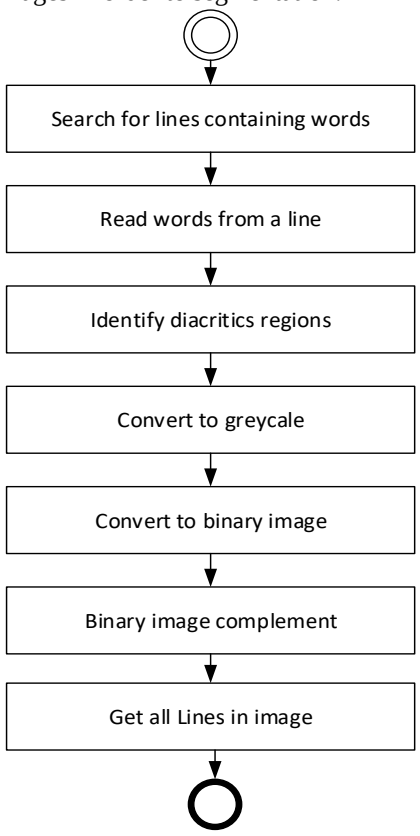

Fig. 3: The Flow of the diacritical detection

VI. IMPROVED K-NEAREST-NEIGHBOR ALGORITHM

The vector-based proposed in this research is the vector which can reflect the essence of classification. It is different from the principal component in principal component analysis (PCA). The principal component is a smaller set of variables by converting the original correlative variable set by means of orthogonal transformation. It may be unable to reflect the essence of Classification, so it is not the vector-based. Similarly, the essential vector is different from prototype vector, the centroid, the kernel function of SVM and the concept of the kernel in knowledge reduction.

I. **The process of the improved KNN algorithm is as follows:**

- Given a test images x;
- Find the K nearest neighbors of x among all the training images,
- Calculate the similarity between the test image vector and the essential vector
- Adjust the number of files in the two file sets compared
- Score the category candidates based on the category of K neighbors
- The similarity of x and each neighbor image vector is the score of the category of the neighbor image.
- Decision function can be defined as follows:

$$\mu_j(x) = \sum_{i}^{k} \mu_j(x) sim(x, x_i)$$

- The distance between the test set x and the training set is determined, and then it is saved into a temporary array

The improved algorithm Vector based kNN first constructs a kNN classifier based on the essential vector to get the *k* candidate classes fast and then constructs another kNN classifier with these *k* classes. So, it reduces the number of training samples through calculation of the first kNN classifier and then uses the second kNN classifier to classify left training samples. The efficiency is therefore increased significantly.

VII. CONCLUSION

Poor recognition rate in most of optical diacritics characters recognition is mainly attributed to failing in segmenting a handwritten text correctly. To overcome this problem, we performs seven operations; it starts with searching the textline height followed by reading words from the line. Then identify the diacritics regions. The segmentation is also applied during this operation by convert the textline into grayscale and binary image. Moreover, we introduced a new method based on k-nearest neighbors (KNN) algorithm to identify diacritics and characters segmentation. KNN is trained to directly predict the diacritic from the textline. Finally, we offer an evaluation discussion on optical diacritics characters recognition.

REFERENCES

1. Abuzaraida, M. A., Zeki, A. M., & Zeki, A. M. (2010). *Segmentation techniques for online Arabic handwriting recognition: a survey*. Paper presented at the Information and Communication Technology for the Muslim World (ICT4M), 2010 International Conference on.
2. Aksoy, S. Introduction to Pattern Recognition.
3. AlOtaibi, S., & Khan, M. B. J. c. (2017). Discovering Semantic and Sentiment Correlations using Short Informal Arabic Language Text. *9*(11), 12.
4. Bennet, J., Ganaprakasam, C., & Kumar, N. (2015). A Hybrid Approach for Gene Selection and Classification using Support Vector Machine. *International Arab Journal of Information Technology (IAJIT), 12*.
5. Di Martino, M., Hernández, G., Fiori, M., & Fernández, A. (2013). A new framework for optimal classifier design. *Pattern Recognition, 46*(8), 2249-2255.
6. El Kessab, B., Daoui, C., Bouikhalene, B., & Salouan, R. (2015). A Comparison between the Performances of Several Distances for Isolated Handwritten Arabic Numerals Recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 8*(6), 9-14.
7. Ibrahim, H. S., Abdou, S. M., & Gheith, M. J. a. p. a. (2015). Sentiment analysis for modern standard Arabic and colloquial.
8. Jayech, K., Mahjoub, M. A., & Amara, N. E. B. J. I. A. J. I. T. (2016). Arabic handwritten word recognition based on dynamic bayesian network. *13*(6B), 1024-1031.
9. Khoshnevisan, B., Bolandnazar, E., Barak, S., Shamshirband, S., Maghsoudlou, H., Altameem, T. A., & Gani, A. (2015). A clustering model based on an evolutionary algorithm for better energy use in crop production. *Stochastic Environmental Research and Risk Assessment, 29*(8), 1921-1935.
10. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. In.
11. Luqman, H., Mahmoud, S. A., & Awaida, S. (2015). Arabic and Farsi Font Recognition: Survey. *International Journal of Pattern Recognition and Artificial Intelligence, 29*(01), 1553002.
12. Ma, G., Xu, Z., Zhang, W., & Li, S. (2015). An enriched K-means clustering method for grouping fractures with meliorated initial centers. *Arabian Journal of Geosciences, 8*(4), 1881-1893.
13. Schantz, H. F. (1982). *The history of OCR, optical character recognition*: Recognition Technologies Users Association Manchester, VT.
14. Singh, D., Khan, M. A., Bansal, A., & Bansal, N. (2015). *An application of SVM in character recognition with chain code.* Paper presented at the Communication, Control and Intelligent Systems (CCIS), 2015.
15. Smith, R. (2007). *An overview of the Tesseract OCR engine.* Paper presented at the Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on.
16. Spera, E., Tegolo, D., & Valenti, C. (2015). *Segmentation and feature extraction in capillaroscopic videos*. Paper presented at the Proceedings of the 16th International Conference on Computer Systems and Technologies.