Journal of Education for Pure Science- University of Thi-Qar
Vol.11, No1 (June, 2021)
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

DOI: http://doi.org/10.32792/utq.jceps.11.01.06

ARABIC LETTERS RECOGNITION by ADOPTING MODIFIED THRESHOLD and CORRELATION APPROACH

Ahmed Hamid Saleh

Northern Technical University / Administrative Technology Collage, Mosul

Received 14/12/2020 Accepted 08/02/2021 Published 10/06/2021

Abstract:

The letters form a front face that can be recognition as a transformative process that transforms the paper document into a fully editable form. The explanatory necessity because of the urgent need associated with the work of the computer, which is based on the processing of the data, shortens the time and limits it based on the accuracy and speed of completion of the computer work as well as clarity by the user.

This paper suggests a new method of letter recognition based on modified the technique of threshold and correlation coefficient, which marks the process of recognition between the 28 letters and letters of the Arabic language. The results obtained in the applied process on the set of these characters in various and multiple sizes(volume letter) (24, 26, 28, 30, 34, 38, 40, 45, 50) and multiple lines (Simplified Arabic, Arial, Traditional Arabic, Tahoma), which effect obtaining high scores when calculating the recognetion ratio of 99.70, this method has the flexibility and speed of other methods used to process this process.

1.Introduction:

The Arabic language is characterized by depth and solid roots, as it is considered one of the first-level languages, which is different from other languages, for example Latin and the languages of the East in Asia and others. With characteristics that take dimensions in terms of the number of words that reach the boundaries of thousands, which differ from the English and French languages that do not exceed 2000 or 3000 words, as well as Hebrew. Due to the practical need, openness to the worlds of the Internet, the overlapping of geographical borders, the space of languages, the need for easy and rapid communication, and the percentage of the population that they speak a language ,so in the latest statistics, which reached 467 million people, thus registering a superiority and progress from the group of Semitic languages, as it is several of the fourth largest languages spoken after Chinese, English and French, and this language in Europe and America represents one of the main languages as it was ranked sixth among other languages beginning with English Spanish and French.[1]

In view of the increasing uses of this language in the world of the Internet and the flexibility it enjoys, the studies of 2017 indicated that it is the most widely used language on the Internet, the specificity of this language and the increasing uses in language processing in the computer and the focus on processing these languages on the basis of regional or geographical specificity and writing in the mother tongue This prompted researchers to work on finding many methods and directions concerned with improving the processing of this language in terms of recognition letters or their composition, so many techniques appeared that relied on optical character recognition that relied absolutely on reading printed or written data that it's recognizes it at a high speed shortens time and effort. In view of the lack of studies have

investigated the topic of applying the distinction between letters, and since the efforts are modest, they need great efforts, especially given the importance of the Arabic language and its computer uses [2].

2- The aim of the research:

This study is considered as a new method to recognition printed Arabic letters (which depend on a color jpeg image) on the basis of which the process of converting or changing the image into a gray image with 2_dimentions and shapes takes from the threshold as a way to produce a binary image, so that after finding the boundaries of the letter drawings are filled with a rag based on an idea that takes from the front-and-back dialectic that we can imagine that the letter is filled with the number (1) and its background with the number (0), and then an estimate is made of the amount of proximity or distance of the letter from Arabic letters by calculating the correlation coefficient for letters with Arabic letters which are taken from percentage error and recognition rate as a tool to recognition between them.

3- OCR (Optical Character Recognition):

Character recognition as a process that tries to convert a paper document into a fully editable form, which can be used in processing and other applications as if it were written by a keyboard.

Srihari and others defined it as containing the letter that is in the form of graphic signs to a symbolic representation that can be dealt with by computer [2].

Character recognition can be classified into two main categories, namely (on line) and (off line). (off line) can be divided into two parts, namely [3]:

1- Magnetic Character Recognition (MCR)

2- Optical Character Recognition (OCR)

The focus of this research is on the recognition of printed characters, which leads to one of the main reasons for recognition of Arabic letters, which is that the characteristics of the Arabic language do not directly allow the implementation of many algorithms used in other languages.

The characteristics of the Arabic language can be summarized [4]:

1- The 28 letters of the Arabic language appear in four forms (isolated, initial, intermediate or final), as the writing style is from right to left, instead of writing from left to right.

2- Some of them have the same shape and differ in the number of points that will be on them, for example the letters (ث, ت, ب) take the same shape but differ in the number of points, one point in the letter ب, two points in the letter ت and three points in the letter ث.

3- Some letters differ in the position of the points they will have, for example the two letters (ن, ب) have the same shape and contain one point, but they differ in the location of the point, one above the baseline (letter ن), and the other below the baseline (letter ب) This distinction can change the meaning of the word.

4- Related works:

The recognition of Arabic letters began in the 1980s, and since then there are many different algorithms, methods, and methodologies that proposed to deal with letters to improve their effective and accurate recognition, along with the recent use of modern techniques for recognizing handwriting by Dehghan & others. [5], Amir & others. [5] [6], Amrouch and others [7], and Ahlam &others. [8].

Hammoud (2008) [9] also presented a method for recognition handwritten Arabic numbers on the basis of extracting the common features between letters based on the Gabor-based features and Support Vector stage, taking 2,100 samples presented by accepting 44 of the book, as it showed average rates for

classification between 99.85% and 97.94%, Three scales, five scales, four scales, and six orientations are used respectively.

Abdul-Azim and Al-Sharif (2008) [10] applied many classification methods to a set of MAD base data, using the Basis Function Support Vector stage in which the classification was performed in two stages. In the first stage, many custom features were extracted from a similar dataset by researchers and used to recognition as the final value of the distinction was 99.48%.

Safaa and Jumana (2017) [11], locating license plate was done based on the intensity detection method and optical character recognition was used with correlation approach and template matching for character recognition, the method was investigated using 40 images, the result that the performance was 86.6%.

5- Finding the threshold:

The following suggested method was used to find the threshold for converting the image into binary form, as follows [12]:

In the proposed method, comprehensively search for the threshold limit that reduces the variance within the (intra-class) by finding the set of weights between the values and as equation (1):

$$\ldots\ldots\ldots\ldots 1\sigma_w^2(t) = w_0(t)\sigma_0^2(t) + w_1(t)\sigma_1^2(t)$$

When w0 and w1 are the probabilities of the values to be separated at the threshold term of t, and σ_0^2, σ_1^2 represent the variance of the two values.

$$\text{As : } w_1(t) = \sum_{i=0}^{L-1} p(i) \text{ and } w_0(t) = \sum_{i=0}^{t-1} p(i) \text{ where } l, t = no.of\ pixel \text{ and } p\ (i) is\ probabilities\ of\ i$$

$$\text{and } \sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = w_0(\mu_0 - \mu_T)^2 + w_1(\mu_1 - \mu_T)^2 = w_0(t)w_1(t)|\mu_0(t) - \mu_1(t)|^2$$

Which represents the probability of W at the class μ and when μ_ (0,1, T) (t) which is: $\mu_T(t) = \frac{\sum_{i-0}^{t-1} ip(i)}{w_T(t)}$, $\mu_1(t) = \frac{\sum_{i-0}^{L-1} ip(i)}{w_1(t)}$, $\mu_0(t) = \sum_{i=0}^{L-1} ip(i)$

The algorithm used to find the threshold term:

1-Calculate histogram and likelihood for each level.

2-Initial values of w0 and μ0.

3-Find all probabilities for the threshold when t = 1,2,3 ……. maximum intensity

3-Update w_i and μ_i.

4-Calculate σ_b^2 (t).

5- Estimate the desired threshold limit corresponding to the highest σ_b^2 (t).

6- Extract the features:

For the purpose of identifying the letter, a correlation coefficient calculation method was adopted to extract the features, since the correlation coefficient is a numerical measure to know whether there is a statistical correlation between two variables or not. There are several types of correlation coefficient, each with its own definition and its own range of usability and characteristics, most of which are used when the value is high (approaching +1.00) is a strong direct relationship, and values close to 0.50 are considered moderate and values below 0.30 are considered weak. A low negative value (close to -1.00) is also a strong inverse relationship, and values close to 0.00 indicate little, if any, relationship [12]. If we assume the two variables x and y to find the correlation coefficient, equation (2) will be used:

$$r = \frac{n(\sum xy)-(\sum x)(y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}} \ldots\ldots\ldots 2$$

The correlation coefficient was calculated for all 28 letters and stored for use as a differentiation database

---

7.The proposed method for recognition Arabic letters:

For the purpose of recognition, will assume that A is the letter to be read from the scanner and store it as a gray image with dimensions Nc × Mc which is represented by the equation(3) is:

$$A= \{xij|0\leq i<Mc, 0\leq j<Nc, xij\in \{0, 1,255\}\} \ldots\ldots\ldots 3$$

This image is converted into binary form using the threshold term and represented by equation (4) is:

$$A= \{xij|0\leq i<Mc, 0\leq j<Nc, xij\in \{0,1\}\} \ldots 4$$

will follow the following steps to recognition the letter (see fig.1 and fig.2).

1- Read the letter A.

2.Calibrate the image to be within the standard dimensions that were adopted during the search, where the image is processed with dimensions [24,44] the best rang was found during the experiment .

3.Converting the image to a gray scale and then converting it to the binary form using the threshold limit.

4-Calculate the edges of an image using the function (Image edge = edge (A).

5.Fill in the form as it is filled with number (1) and its background with number (0).

6.Calculating the correlation coefficient for the letter and comparing it with the correlation coefficients for the characters stored in the database (fig.3) by using feautre extraction techniques and displaying the result.

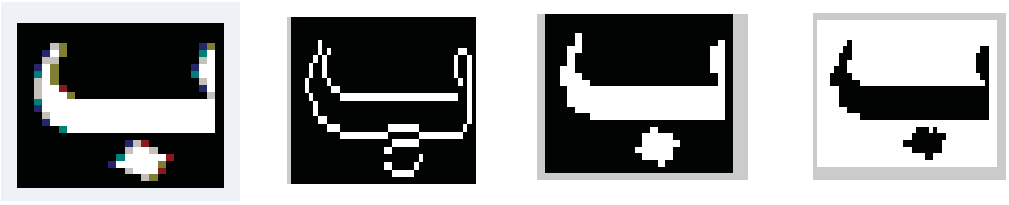

Fig.1: a- The letter to be recognition
b- Applying the threshold and edges to Image No 1
c-the letter after filling the edges
d- The letter in binary form

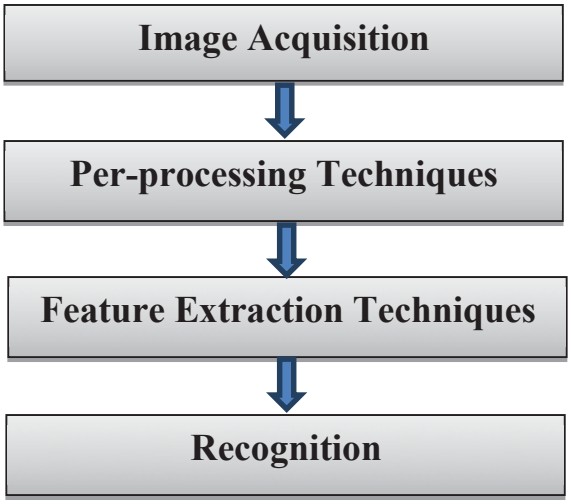

Fig.2:ecognition the Arabic letters

fig.3: template of the character using a database for recognition

8 -Results: -

For the purpose of the experiment, the proposed algorithm was applied to the 28 basic letters of the language and in the (Simplified Arabic) font of size 24, as the correlation parameters for it were found, stored and used as a database later to differentiate the letters.

The proposed algorithm was tested on the 28 characters with a number of sizes (24, 26, 28, 30, 34, 36, 38, 40, 45, 50) and in the (Simplified Arabic) font, as the results were identical as the percentage error that was calculated Using equation (5)

$$p.e.=\left|\frac{c.c.-\overline{c.c.}}{c.c.}\right| * 100\% \quad \ldots\ldots\ldots 5$$

So that

c.c = the value of the original letter's correlation coefficient

$(c.c)^{-}$ = the value of the correlation coefficient for the character to be recognition.

So p.e = 0 and the recreation rate ratio was calculated using equation (6):

$$R.R.= (100\text{-}P.\ E)\ *100\% \quad \ldots\ldots\ldots 6$$

Since its value was 100% for most of the letters, an experiment was also conducted using a number of fonts (Simplified Arabic, Arial, Tahoma, Traditional Arabic) with a size of 28 to measure the efficiency of the proposed algorithm, if the results are as per Table (1).

Table (1) The percentage of discrimination for Arabic letters by adopting a number of fonts and size 28

character	R.R for simplified arabic	R.R for Arial	R.R for tahoma	R.R. for traditional arabic
ا	100%	100%	100%	100%
ب	100%	100%	100%	100%
ت	100%	100%	100%	100%
ث	100%	100%	100%	100%
ج	100%	100%	100%	100%
ح	100%	100%	100%	100%
خ	100%	100%	100%	100%
د	100%	100%	100%	100%
ذ	100%	100%	100%	100%
ر	100%	100%	97.24	100%
ز	100%	100%	97.63	100%
س	100%	100%	100%	100%
ش	100%	100%	100%	100%
ص	100%	100%	100%	100%
ض	100%	100%	100%	100%
ط	100%	100%	100%	100%
ظ	100%	100%	100%	100%
ع	100%	100%	97.52%	100%
غ	100%	100%	100%	100%
ف	100%	100%	100%	100%
ق	100%	100%	98.61%	100%
ك	100%	100%	100%	100%
ل	100%	100%	100%	100%
م	100%	100%	100%	100%
ن	100%	100%	100%	100%
ه	100%	100%	97.41	100%
و	100%	100%	100%	100%
ي	100%	100%	100%	100%
Average	100%	100%	99.70	100%

Through the results, we note the stability and flexibility of the proposed algorithm in recognizing multiple shapes and sizes of letters at high speed, which reduces time and effort.

9- Conclusions:

The adoption of the correlation coefficient method, which works to find the proximity of two variables to each other or not, gave high results in character recognition, which was adopted by many researchers in recognition Chinese and English characters in addition to numbers, and through the results reached by the study we note and when adopting the algorithm The proposed research gave a high possibility to recognition Arabic letters with less time and effort, and gave a high amount of discrimination if the discrimination ratio

reached RR = 99.70, which is a very high percentage in discrimination compared to previous algorithms in this field.

10-Resource:

[1] Cheriet, M., N. Kharma., (2007). Character Recognition Systems: A Guide for Students and Practitioners, Wiley.

[2] Srihari, S.N., A. Shekhawat, and S.W. Lam. (2003), Optical character recognition (OCR).

[3]Ahmed, T., Mohammed, R. and Abdelhay, S. (2014). Investigating of preprocessing techniques andnovel features in recognition of handwritten Arabic characters. Lect. Notes Artif Int., 2014; 264-276.

[4] Rawia I. O. Ahmedو Mohamed E. M. Musa. (2016). Preprocessing Phase for Offline Arabic Handwritten Character Recognition.

[5]Dehghan, M., Faezl, K., Ahmadi, M. and Shridhar, M. (2000). Off-line unconstrained Farsihandwritten word recognition using fuzzy vector quantization and hidden Markov word models .In: IEEE 2000 Proceedings of the 15th International Conference on Pattern Recognition 3-7September 2000; Barcelona, Spain: IEEE. pp. 351–354.

[6]Amir, M, Karim, F. and Abolfazl, T. (2002). Feature extraction with wavelet transform forrecognition of isolated handwritten Farsi/Arabic characters and numerals. In: IEEE 2002International Conference on Digital Signal Processing; 1-3 July 2002; Santorini, Greece,IEEE. pp. 923 – 926 .

[7]Amrouch, M., Elyassa, M., Rachidi, A. and Mammass, D. (2008). Off-Line Arabic handwrittencharacters recognition based on a hidden Markov models. In: IEEE 2008 International Conferenceon Image and Signal Processing; 1-3 Jul; Cherbourg, France: IEEE. pp. 447 – 454 .

[8]Ahlam, M., Akram, H., Khalid, S. and Hamid T.,(2014). Using HMM toolkit (HTK) for recognitionof Arabic manuscripts characters. In IEEE 2014 International Conference on MultimediaComputing and Systems; 14-16 April 2014; Marrakech, Morocco: IEEE. pp. 475 – 479.

[9] Mahmoud SA. 2008. Arabic (Indian) handwritten digits' recognition using Gabor-based features. In: Innovations in Information Technology, 2008. IIT 2008.International conference. Piscataway: IEEE, 683–687.

[10] Abdul-Azim, El-Sherif E. 2008. Arabic handwritten digit recognition. International Journal of Document Analysis and Recognition,11(3):127–141.

[11] Safaa S.O. Jumana A.J. Iraqi car license plate recognition using ocr. annual conference on new trends in information and communication technology application. Iraq.2017.

[12] Otsu, N., (1975), A threshold selection method from gray-level histograms, Automatic.