Efficient Approach to Segment Ligatures and Open Characters in Offline Arabic text

Zerdoumi Saber[1], Aznul Qalid Md Sabri [2], Amirrudin Kamsin, Saqib Hakak [3]

Abstract-This paper researches offline Arabic handwriting recognition. It introduces a new approach to segmentation ligature and open Arabic character based on the structural perspective dealing with sub-words/words, including dots to recognize individual letters. Segmentation approaches that have been integrated into the recognition phase have the capability to deal with ligatures and closed characters issues. This complex problem is due to the cursive writing nature of the Arabic language. This paper also develops an Arabic character algorithm on segmented, pixel based and centre reign (CR) to recognize the letters. The evaluation results are stated in the IFN/ENIT and IAM database which indicate the recognition rate and the effectiveness of our system.

Index Terms- Arabic Cursive, Pattern Recognition, Image Analysis, Handwriting, Recognition System**.**

I. INTRODUCTION

The Arabic language is one of the world's main languages with over 420 million native speakers in various Arab countries. It also holds a huge influence on ethnic heritage and exactions of the major languages in non-Arab countries such as the Chad and the Central African Republic. It is also a minority Arabic language used in different countries including Afghanistan, Nigeria and Iran. Referring to 1974, the Arabic language was a dominant spoken language in addition to the six main languages listed by the United Nations as official languages side by side with English, Chinese, Spanish, Arabic, French, Russian [1]. With over 1.6 billion Muslims around the globe such as Indonesia, Malaysia, Pakistan, India and Tanzania, the Arabic language is a second language in liturgical and intellectual fields. This drives the demand for learning to read and use Arabic.

The introduction of novel broadcasting information, tools, and media in the last two decades has increased the demand for comprehensive Arabic, managing data and understanding the information captured by various devices. The recent development in social media and the Internet has produced an overwhelming stream of data in both structured and unstructured formats [2].

[1,2,3]University Malaya, Department of Computer Science and Information Technology, Malaysia

One pertinent information type is images. The principal role is to handle the Arabic language by extracting texts from images. However, variance in writings was stored as complex patterns and emerge as pictures or video. For a general text recognition system, the researchers have been able to develop many text recognition systems [3] with high recognition rates such as offline Arabic handwriting and it deals with segmentation characters. The handwriting recognition systems are commonly used to comprehend text from images. Early work began by applying conventional Optical Character Recognition (OCR) techniques, and the results are passed to special search engines to search for words [4]. Unfortunately, the research is still at the early stage and facing many challenges [5]. There are several efforts to address Arabic handwritten word recognition for using certain classifiers such as Support Vector Machines (SVM), HMM and Recurrent Neural Networks (RNN) [6], Transparent Neural Networks [7] and Neural Networks (NN). Mohammad and Sabri introduced a recognition system for Arabic handwriting using a syntactic and structural pattern attributes. The work has attempted different slant angles of various components of single text line where highlighted. Moreover, they invented, *"novel design segmentation algorithm which is integrated into the recognition phase"* [7].

The rest of this paper is structured as follows. Section 2 presents the related definitions and characteristics of the Arabic language. Section 3 provides an overview of the recognition system. Section 4 proposes an extraction system based on segmentation region; Section 6 provides some directions for future works in this research domain.

II. RELATED WORK

This section reviews studies that focus on different state-of-the-art techniques and approaches for image segmentation and recognition processes that deal with the Arabic characters. This is to provide a concise summary of the research that has already been done in the field.

Segmentation is where sub-methods deals with only pattern image recognition. The approaches applied for Arabic handwriting include overlapping, close characters issues composite character and touching character. Each of these methods deals with Arabic text line, starting with digitisation, noise removal, skew detection, and correction are required in the preprocessing stage. To recognise text in an image, there is high demand for segmentation; the whole text is converted into lines text and then single words. It is then converted into individual characters by performing the vertical, horizontal projection [8]. Unfortunately, research's is still immature and

facing different challenges[5]. For example, developing Arabic OCR systems is a difficult task because of three reasons. Firstly, the Arabic cursive is printed from the right side to left side [3]. Secondly, the Arabic characters have different shapes, which can change according to whether the alphabet shape is connected from the beginning to the middle and from the middle to the end. Lastly, the Arabic script has diverse Arabic fonts on printing machines and recognition systems. This may suggest that a specific approach designed for Arabic writing recognition should be established.

TABLE1
ARABIC ALPHABET AND ITS APPEARANCE IN DIFFERENT POSITION (BEGINNING – MODEL –AND END):

Isolated alphabets	Right sound	Beginning (B)	Middle (M)	End (E)	Pronunciation when character is combined with vowels (a, o and i)
Alfeأ	A	أ	أ	أ	Aa, Ao,Ai
Baa ب	B	بـ	ـبـ	ـب	Ba, BO, BI
Taaت	T	تـ	ـتـ	ـت	TA,TO,TI
Thaaث	TH	ثـ	ـثـ	ـث	THA,THO,THI
Haaح	H	حـ	ـحـ	ـح	HA,HO,HI
Jimeج	Dj	جـ	ـجـ	ـج	DJA,DJO,DJI
Khaaخ	Kh	خـ	ـخـ	ـخ	KHA,KHO,KHI
Dale د	D	د	ـد	ـد	DA,DO,DI
Dthalذ	DH	ذ	ـذ	ـذ	DHA,DHO,DHI
Raaر	R	ر	ـر	ـر	RA,RO,RI
Zaayز	Z	ز	ـز	ـز	ZA,ZO,ZI
Sineس	S	سـ	ـسـ	ـس	SA,SO,SI
Shineش	Sh	شـ	ـشـ	ـشـ	SHA,SHO,SHI
Sadeص	S	صـ	ـصـ	ـص	SA,SO,SI
Daadض	D	ضـ	ـضـ	ـض	DA,DO,DI
Taa'dط	T	طـ	ـطـ	ـط	TA,TO,TI
Zthadظ	Z	ظـ	ـظـ	ـظ	ZA,ZO,Zi
Aineع	A	عـ	ـعـ	ـع	AA,AO,AI
Ghayneغ	Gh	غـ	ـغـ	ـغ	GHA,GHO,GHI
Faaف	F	فـ	ـفـ	ـف	FA,FO,Fi
Quafق	Q	قـ	ـقـ	ـق	QA,QO,QI
Kaafeك	K	كـ	ـكـ	ـك	KA,KO,KI
Lame ل	L	لـ	ـلـ	ـل	LA,LO,LI
Mime م	M	مـ	ـمـ	ـم	MA,MO,MI
Nuneن	N	نـ	ـنـ	ـن	NA,NO,NI
Haaه	H	هـ	ـهـ	ـه	HA,HO,HI
Waawو	W	و	ـو	ـو	WA,WO,WI
Yaaي	Y	يـ	ـيـ	ـي	YA,Yo,Yi

2.1 Arabic Characters Orthography and pronunciation

The Arabic language is like the Latin language with alphabets named Hourouf- (حروفpossesses 28 characters), but it has a unique Morphological form. It has its own specific language orthography and morphology of charters. Each character has two to four different shapes, and the selection of shapes depends on connecting characters within several words or sub-words. The shapes correspond to the positions. Namely, the single alphabets that can be changed according to it is connectivity from "the beginning" to" the middle" and from "the middle to the end" (sub-) words and in isolation of (sub-) words. Table 1 shows each single pattern. The shape can be changed according to its sides and attached to its neighboring sides on each one and is connected to its neighbourhood on each news. Several characters have an initial or middle changing shapes. They are simply written only at the end of (sub-) words.

2.2 the process of arabic cursive recognition system based on extraction features

Illustrating the comprehensive study based on image recognition system, our discussion is to give a clear picture for each sequence step in recognising features from input source to final stage crossing, preprocessing, segmentation, classification, and reporting case studies on recognition rate using different comparisons between Arabic character recognition systems. This section is split into two portions. The first character deals with a pre- processing stage that integrates images into their processing levels.

2.3 Classification Pattern Recognition Processes Using Approaches

Recognition processes are split into different classes. Each category is spread through step sequences. To better comprehend their characteristics, numerous classes of recognition processes are used. This classification is important because large-scale resources can be used and extracted in the recognition stage. The categorisation is founded on five aspects, namely, (i) pre-segmentation, (ii) segmentation, (iii) feature extraction, and (IV) and (v) post-processing approaches. Fig1 shows main recognition steps for Arabic recognition systems.

(1) Segmentation refers to splitting images, detecting the contour of each word, and classify it into sub-words collected via single processing[9, 10]. The second signal views the changes in the system states in sequential order. This analysis makes detections of local least of the upper contour in a structured format. Most classical segmentation approaches are generated by using OCR based on separate horizontally overlapping Arabic word/sub-word separation[11].

(2) Feature extraction refers to the combination of comprehensive approaches to character recognition and character segmentation, such as contour tracing, contour analysis, and sub-word detection[12, 13]. Feature extraction is the transferring of image contents to an understandable text with counting absorption complementary errors caused by system previously archived.

(3) Classification is the most significant aspect of the Arabic recognition system. It refers to the process of associating a shape with a corresponding character from large values and large datasets with various handwriting types.

The following representation is proposed on the basis of the above-mentioned designations and statement in the analysis of image recognition. The image identification system is a set of techniques and approaches depending on their technologies required for each footprint in the flow chart sequence with some addition, such as machine learning largely hidden values from specific datasets that require varied and complex cursive writing.

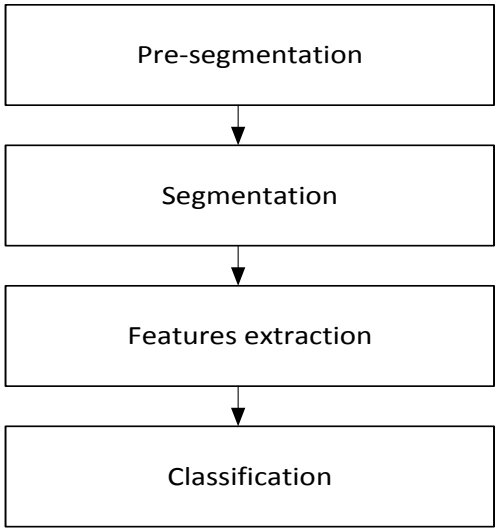

Fig.1. Diagram of the four main recognition steps for Arabic recognition systems.

III. PROPOSED APPROACH

After the preprocessing stage, the binarization of the image into small modules is performed by splitting the image into specific text lines. This segmentation is done via vertical and horizontal projection of pixels where a profile projection is an overlooking histogram with the cumulative quantity of black pixels within the same line. Through consecutive dual summits in the horizontal profile, the boundary text can be performed between two projection text lines.

The recognition, ligatures and close character are investigated to segment closed characters. Using, foreground pixels defined as $\sum_{vi}^{n} Pi$ sum of pixels in each column within thinned word image, the sum of columns were considered as word. The segmentation is based on pixel value of centre reign (CR) where count start is 1 or 0. This preliminary phase is useful to trace alphabets boundaries by alphabets then recognise the whole word. Arabic character segmentation has its own specification which led to more effort to solve the complexity faced in ligature and overlapping. In the ligatures/union of characters, segmentation is performed using deference character as template matching to recognise letters. However, the segmentation of open characters (loop/semi-loop) cannot distinguish characteristics in segmentation steps due to the presence of within-letter-ligatures. Therefore, additional features are required for holder and truing. A natural network is needed to deal with the problem of segmentation and to detect precise character boundaries.

First of all, ligatures segmentation is done by a middling centre of gravity (CR) of the character before taking the alphabet which is calculated by minimum distance less than a specified threshold to merge them into one segmentation column. The threshold is the minimum horizontal distance between successive CR that could not accommodate a character and re-sets the value to re-experimental loops.
This phase ensures valid ligature segmentation.

For the case segmentation of open characters. In the preceding step, ligature analysis confirms correct segmentation of closed characters. Nevertheless, open characters are still not well segmented. For illustration, an open characters 'ي' and 'ن' are segmented into two parts. Both parts do not comprise loop/semi-loop in segmentation processes.
To handle segmentation in the open characters, each segmented letter is over demanded by using extra features and neural confidence.

Therefore, foreground pixels in each segmented word are counted. Such segmentation columns are identified and extracted. The results are shown in the flowing system that has been testing for different handwriting and can extract text with over 98% recognition rate

PROPOSED ALGORTHIM LIGATURES CHARACTER SEGMENTATION

Processing

1. **Apply thinned filter for the word image**
2. **The sum of the foreground pixels were considered as a word.**
3. **Indicate pixel value of centre reign (CR), counting association (0 and 1) with CR column**
4. **From 1.2.3 trace letter boundary.**

Segmentation:

5. **Normalise centre gravity of the reign**
6. **Calculate minimum distance in threshold image**
7. **Merge six in segmentation column**
8. **Apply horizontal distance between successive CR.**
9. **Segment character**
10. **Move to 12.**
11. **If not segmenting a ligature character go to step 5.**
12. **End Algorithm.**

PROPOSED ALGORTHIM OPEN/CLOSED CHARACTER SEGMENTATION

Processing

1. **Apply thinned filtered for the word image.**
2. **The sum for foreground pixels were considered as a word.**
3. **Indicate pixel value of centre reign (CR), counting association (0 and 1) with CR column**

4 From 1.2.3 trace letter boundary.
Segmentation:

5 Apply loop/semi-loop for "LIGATURES CHARACTER SEGMENTATION" algorithm
6 Additional features demand via neural confidence.
7 Find approximate /suitable open characters for matching.
8 End Algorithm.

Fig. 2 Develop Arabic word recognition system based on segment ligatures and closed characters in offline Arabic text.

Previous proposed Algorithms deal with the issue of segmentation ligatures character and open, character segmentation by giving greatest association to recognise Arabic characters including the complexity related to the cursive style. Further prediction supplements reliable dataset. Fig 2 demonstrate the Development of the Arabic word recognition framework based on segment ligatures and closed characters in offline Arabic text.

Few datasets have strong association related to this issue such as the IFN/ENIT database[14].

IV. Evaluation and Analysis

Recognition based on segment ligatures and open characters including segmentation region has given accuracy results of over 98% recognition rate as displayed in Tables 2 which has been tested for different handwriting datasets IFN/ENIT - database Arabic OCR handwritten[14] with 2% over-segmentation. The measurement segmentation performance tests for each experiment test targets the contained open Arabic letters (ي,ح,ق.ة.ن) and the results are presented in Table 2. The Column was named "Quasi" and quasi-word i.e. (a word that comprises poorer character or noise). With typical deviation confirmation, the quantity of quasi-word and over-segmentation is abridged momentously. However, it should be noted that average deviation verification is complete by removing and distinguish characters within the baseline. Thus, this approach is disposed to open characters segmentation. In both tests, the segmentation is not initiated but if we loop it, vivification segmentation will increase the rate. By repeating the verification process only twice, segmentation is increased by 2%. Word segmentation performance with incorporation standard deviation verification is 93.9%.

The classification phase has accurateness of 82% via cross validation.

Table II
Performance of whole recognition

Data set type	Without standard deviation			With standard deviation		
	Correct	Over	Quasi	Correct	Over	Quasi
IFN/ENIT	98%	2%	21%	100%	0%	0%
QUWI	99%	1%	9%	100%	0%	0.8%
Own Feature	97%	3%	33%	99.4%	0.6%	0%

This accuracy is reached by using 24 features for generating the decision tree and using the normal and bold image of the letter as the training set.

Performance of segmentation characters

Table III
Performance of open Characters segmentation.

Dataset type	*Open Characters tested*	*Correct*	*Over*	*Quasi*
IFN/ENIT	ن	82%	13%	5%
QUWI	ي	79%	17%	3%
Own	ق	91%	2%	6%
Feature	ح	92%	2%	6%
	ه	89%	4%	5%

The key factor for performing a good recognition body for segmentation characters is to split all letters into sub-classes which is ten similar letter morphology forms and eight classes representing different shapes. This classification helps to increase the recognition rate by reducing mismatching and error recognition. This produced accuracy over 82%. Before modelling this classification, we had very poor results with only 46% accuracy. This increased to 79% after incorporating interest point features and a number of the loop such as the letter (ي). This increased to 82% after compressing the amount of target class into 23 letter shapes. This main body-classes squeezing is done by connection similar linked letter procedures from a letter into one class. For illustration, ـجـ and ـحـ have two classes before compressing but joined into one class after compressing.

V. Comparative Study

The comparative results uses different related systems and Benchmarking IFN/ENIT as a database. This is a database of handwritten Tunisian city names. It is divided into type of

model and recognition rate. Table 3 summarises this comparison.

TABLE IV
COMPARISON RECOGNITION RESULTS WITH A RATE OVER 98% COMPARED WITH OTHERS SYSTEMS.

Author	Model	Recognition
Proposed	-	98%
[15]	ANN using CEDAR dataset	95.27%
[16]	ANN using IAM dataset	85.36%

Our proposal model gives better accuracy results with over 98% recognition rate for overlapping and open letters compared with existing methods that used the CEDAR dataset for 95.27% accuracy and using IAMdataset to produce 85.36% accuracy.

VI. CONCLUSION AND FUTURE WORK

This paper begins with a study of the Arabic language, including its recognition challenges, methods applied for letter recognition in general description, morphology and orthography. It compared several methods for classifying data as used by the different recognition systems. The paper proposed a new algorithm for dealing with ligatures and open characters in Arabic text. Different resources were applied to extract a character from a features format to readable text. The aim is to improve the Arabic recognition system in the pattern recognition field (APR), and its influences in both speech recognition and natural language applied in Arabic. In the future, we are planning to carry out an experiment and propose a model can test various documents as parameters including Othman script as written in rhe Holy Quran. The model will also be tested using different module image size, after a number of generations and training data to determine the letters and to reduce the computational duration time.

REFERENCES

[1] Al-Badr, B. and S.A. Mahmoud, Survey and bibliography of Arabic optical text recognition. Signal processing, 1995. 41(1): p. 49-77. https://doi.org/10.1016/0165-1684(94)00090-M

[2] Zhang, H., et al., Text extraction from natural scene image: A survey. Neurocomputing, 2013. 122: p. 310-323. https://doi.org/10.1016/j.neucom.2013.05.037

[3] Lutf, M., et al., Arabic font recognition based on diacritics features. Pattern Recognition, 2014. 47(2): p. 672-684. https://doi.org/10.1016/j.patcog.2013.07.015

[4] Davis, B., R. Clawson, and W. Barrett, Flexible Computer Assisted Transcription of Historical Documents Through Subword Spotting.

[5] Saber, S., et al. Performance Evaluation of Arabic Optical Character Recognition Engines for Noisy Inputs. in The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt. 2016. Springer.

[6] Azizi, N., et al., Using diversity in classifier set selection for arabic handwritten recognition, in Multiple Classifier Systems2010, Springer. p. 235-244. https://doi.org/10.1007/978-3-642-12127-2_24

[7] Parvez, M.T. and S.A. Mahmoud, Arabic handwriting recognition using structural and syntactic pattern attributes. Pattern Recognition, 2013. 46(1): p. 141-154. https://doi.org/10.1016/j.patcog.2012.07.012

[8] Mozaffari, S., et al., Two-stage lexicon reduction for offline Arabic handwritten word recognition. International Journal of Pattern Recognition and Artificial Intelligence, 2008. 22(07): p. 1323-1341 https://doi.org/10.1142/S0218001408006843

[9] Olivier, G., et al. Segmentation and coding of Arabic handwritten words. in Pattern Recognition, 1996., Proceedings of the 13th International Conference on. 1996. IEEE.

[10] Alginahi, Y.M., A survey on Arabic character segmentation. International Journal on Document Analysis and Recognition (IJDAR), 2013. 16(2): p. 105-126. https://doi.org/10.1007/s10032-012-0188-6

[11] Cheung, A., M. Bennamoun, and N.W. Bergmann, An Arabic optical character recognition system using recognition-based segmentation. Pattern recognition, 2001. 34(2): p. 215-233. https://doi.org/10.1016/S0031-3203(99)00227-7

[12] Kavianifar, M. and A. Amin. Preprocessing and structural feature extraction for a multi-fonts Arabic/Persian OCR. in Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on. 1999. IEEE.

[13] Azmi, A.N., D. Nasien, and S.M. Shamsuddin, A review on handwritten character and numeral recognition for Roman, Arabic, Chinese and Indian scripts. arXiv preprint arXiv:1308.4902, 2013.

[14] Abandah, G.A., F.T. Jamour, and E.A. Qaralleh, Recognizing handwritten Arabic words using grapheme segmentation and recurrent neural networks. International Journal on Document Analysis and Recognition (IJDAR), 2014. 17(3): p. 275-291. https://doi.org/10.1007/s10032-014-0218-7

[15] Cheng, C.K., et al. Enhancing neural confidence-based segmentation for cursive handwriting recognition. in 5th International Conference on Simulated Evolution and Learning, Busan, Korea, SWA-8. 2004.

[16] Rehman, A., et al. Performance analysis of segmentation approach for cursive handwriting on benchmark database. in 2009 IEEE/ACS International Conference on Computer Systems and Applications. 2009. IEEE. https://doi.org/10.1109/AICCSA.2009.5069335