A SURVEY ON TECHNIQUES USED IN ARABIC HANDWRITTEN CHARACTER RECOGNITION

V.M. Ashiq, E.J. Thomson Fredrik*

Abstract

Arabic handwritten character recognition is one of the challenging tasks in the field of Natural Language Processing. Currently there are so many English recognition methods. But due to the diversity in Arabic character's position and shapes, there are only few Natural Language Processing metods available for Arabic Language. This research paper is an attempt to study various algorithms used for Arabic handwritten character recognition. Latin script OCR is a well-researched area. Arabic script OCR is an emerging area of intense research that follows few reasons: Firstly, around 200 million people in the world use Arabic as their first language. Secondly, around 1.6 billion people follow Islam wherein it is compulsory to recite the religious scripture which was revealed in Arabic. Thirdly, after achieving considerable success in Latin Text OCR, researchers have now focused on extending their prowess to the more challenging Arabic text OCR. Hand written character recognition using many methods like Support Vector Machine analysis algorithm, K Nearest algorithm, Bioinspired alogorithm like artificial neural networks and convolutional neural networks etc. [1,2,3] .In this study, preprocessing, feature extraction and post processing methods are focused.

Keywords: OCR,Character data base recognition,Neural Networks.

I INTRODUCTION

This paper focusses on the basic concept of OCR and also its importance. An overview of the basic Document Structures: geometric and logical, has been presented. There are so many fields in various fields like digit recognition, Bank cheque analysis [4-11],Office automation [12-15], Document processing[16], document content based retrieval[17], Signature verification [7][11],Postal code recognition[4,5,7-11] and digital character identification system.The algorithms used for word recognition have been divided into three classes, namely holistic approach, analytic approach and feature matching approach. The offline Arabic Handwriting Recognition problem has also been defined.

The majority of proposed methods for dealing with Arabic characters involve dimensionality reduction of retrieved characteristics. In some procedures, the original features should always be translated to another domain (like Principal Component Analysis). To get around this limitation, feature-selection strategies are used, even if they aren't always effective in limiting back the most useful traits. In feature-selection algorithms, a number of swarming techniques have been developed to improve the process of discovering the far more relevant features [18].

I.1 OCR (Optical Character Recognition)

Optical Character Recognition (OCR) is the process of converting the spatial representation of text e.g. a scanned document, into its symbolic representation e.g. ASCII Characters. An automated recognition of handwritten text is implied. This is useful in a number of constantly emerging applications. It is useful for processing scanned documents, acquiring data/commands through hand-held devices, verification of writers' identities, etc.

I.1.1 History of OCR

Research on OCR has come a long way from its humble beginnings. Since then OCR research has gathered much

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

momentum and has absorbed a significant amount of science. Several important surveys have been noticed . A good number of relevant papers have been published in journals and conference proceedings .Several books have been published on relevant topics. As faster processing and bigger storage became increasingly available and affordable, document processing systems became more robust and accurate. The automatic extraction of useful information from various on-line data sources in various languages has been gaining momentum.

I.2 Basic Document Processing Model

There are two main types of written documents available, viz. machine-printed text and handwritten text. This research concentrates on ways to improve automated recognition of text with special reference to Arabic text. Automatic document processing involves recognizing text and/or images, followed by extracting the desired information, in a format acceptable for humans. The main steps in Processing of a document is mentioned below.

A document has mainly classified into two structures: a) Geometric (layout) structure and b) Logical Structure. The former represents the objects of the document and the connections amongst these objects, on the basis of presentation. The latter represents the objects in the document and the connection among these objects as per the classification done by a human.

Document processing proceeds in two phases

a) document analysis means to extract the geometric structure of documrnt/image.and

b) document understanding means the mapping of the geometric structure into its related logical structure. After having captured the logical structure, techniques like Artificial Intelligence can attempt to decode its meaning. Figure 1, shows the relationships between the geometric structure, logical structure, document analysis and document understanding.

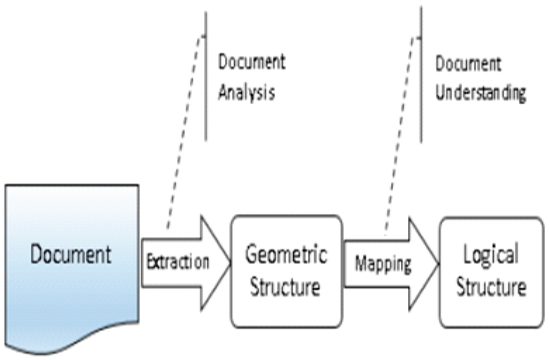

Figure 1: Basic Document Processing Model

I.3 Strategies for Recognition

The various categories of word recognition algorithms are:

- Holistic Approach
- Analytic Approach
- Feature Sequencing Matching Approach

The holistic approach attempts to recognize the entire word by utilizing shape features extracted from the word image. Holistic methods are feasible when a small number of words are to be recognized.

The analytic approach involves segmenting the word image into primitive components, like characters, or Part-of-Arabic-Words (PAW's). The practice of segmentation of a word into characters before recognition is called external character segmentation, while the practice of concurrent execution of segmentation and recognition is called internal character segmentation.

The feature sequence matching approach involves extracting features sequentially and deriving word identity from this sequence. Many pattern recognition approaches are being used in Arabic Word Recognition research as of now. These include Artificial Neural Networks (winner of ICDAR 2013 OCR Competition), the versatile Hidden Markov Models, Support Vector Machines (winner of ICDAR 2015

Writer Identification Competition), Dynamic Bayesian Networks, etc.

II. PROBLEM STATEMENT

Recognition of Arabic handwritten words poses a problem which needs to be addressed with regard to office automation, versatile e-search, versatile e-understanding of online literature, as well as many other applications. The pattern of Arabic characters is not fixed. However, their geometrical features are fixed. Arabic characters differ in their shapes between writers, but their geometrical features are always similar. An important difference between Arabic words and their Latin counterparts is the existence of dots and diacritics (pronunciation marks). These additional impressions differentiate between characters with similar geometry. Another difference is the presence of multiple baselines in Arabic text as against a single baseline in Latin text. Further the baseline may be altogether absent in certain compact Arabic writing styles.

These factors make OCR for Arabic handwritten text difficult. This research introduces novel techniques for improving performance of both conventional OCR Engines and handwritten Arabic text OCR Engines. Arabic handwriting OCR involves several factors summarized below.

- Cursive Arabic handwritten characters are different from their machine printed counterparts.
- Arabic writing is different from English writing in many ways.
- Off-line recognition is different from on-line recognition in certain respects.

III. LITERATURE SURVEY

In [19], Azuraliza, A.B., Siti Rohaidah, A., Nurhafizah Moziyana, M.Y., Yaakub, M.R.,(2017) proposed an unique FS technique for sentiment analysis based on Ant-Colony-Optimization, as according them (ACO). Researchers evaluated the effectiveness of this suggested approach using a KNN classification using customer feedback datasets. The results obtained were compared using Information-Gain (IG), Genetic-Algorithm (GA), and Rough-Set-Attribute-Reduction (RSAR). The researchers proposed the most precise findings, with an improved precision of 0.914.

In [20] Mudhsh MA, Almodfer R (2017) introduced an Alphanumeric very Deep-Neural-Network as a strategy for identifying handwritten Arabic numerals and characters. A classification model was created using 13 convolutional, 2 max pooling layers, and 3 fully connected layers. Two leveling approaches, Augmentation and Dropout, were employed to minimise the number of parameters. The AD-Base dataset (a collection of Handwritten Arabic values from 0-9) and the HACDB dataset were used in the testing (a dataset of Characters with Arabic Handwritten). The ADBase dataset has a 99.67 percent accuracy rate, whereas HACDB dataset has a 97.42 percent accuracy rate.

In[21] Younis K (2018) created a CNN that can recognise Arabic handwritten characters. In their proposed CNN, three convolutional layers were suggested, and including fully - connected layers. Testing results indicated that the CNN could achieve 94.8 percent and 94.9 percent accuracy, however, employing the AHCD and AIA9K database.

For detecting Arabic handwritten characters In [3] Naseem Alrobah and Saleh Albahl, 2021 proposed a hybrid approach. They created a convolutional neural network using a hybrid model support vector machine and eXreme gradient boosting classifiers using an Arabic dataset named Hijja. This algorithm has a 96.3 percent recognition rate. Arabic Handwritten Character Recognition (AHCR) has recently become an important topic in pattern recognition and computer vision. Different Machine Learning (ML)

techniques, such as Support Vector Machines (SVM) and Artificial Neural Networks, significantly improved AHCR technology (ANN).

The researchers presented a novel FS approach based on Ant-Colony-Optimization for Sentiment-Analysis in [19]. (ACO). Researchers investigated the effectiveness of this recommended technique using a KNN classifier using customer feedback datasets. The findings were compared using Information-Gain (IG), Genetic-Algorithm (GA), and Rough-Set-Attribute-Reduction (RSAR). The researchers proposed the most reliable data, with an improved accuracy of 0.914.

In [20], the researchers presented an Alphanumeric very Deep-Neural-Network for recognising handwritten Arabic numerals and characters. A classification model was constructed using 13 convolutional layers, 2 max pooling layers, and 3 totally connected layers. Two normalisation approaches, Augmentation and Dropout, were utilised to minimise the number of parameters. The AD-Base dataset (a collection of Handwritten Arabic values from 0-9) and the HACDB dataset were used in the testing (a dataset of Characters with Arabic Handwritten). The ADBase dataset has a 99.67 percent accuracy rate, whereas the HACDB dataset has a 97.42 percent average accuracy.

The researchers from [21] created a CNN that can detect Arabic handwritten characters. In their proposed CNN, three convolutional layers were offered, along with totally linked layers. Testing results demonstrated that the CNN could achieve 94.8 percent and 94.9 percent accuracy, respectively, utilising the AHCD and AIA9K datasets. To improve the findings, the researchers in [22] combined Simulated-Annealing with a hybrid based FS. They used 11 regressions and 29 classification datasets to test the novel method and compare it to existing methods. These are all favourable results.

IV. METHODOLOGIES

A basic foundation for Arabic text handwritten recognition systems are: Preprocessing, representation, segmentation, feature extraction, and recognition are all common components of recognition algorithms. Figure 2 illustrates the components of the proposed framework, which are organised as in typical applications.

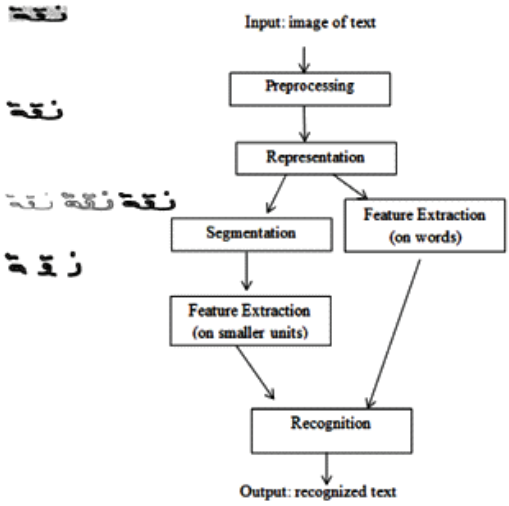

Figure 2: Methodologies

(i) Preprocessing

The raw data is subjected to a variety of preliminary processing steps, depending on the kind of capture, in order to make it useable in the descriptive phases of character analysis.

1) Noise reduction,
2) Normalization, and
3) Connected components extraction all are the elements of the preprocessing step of handwritten Arabic script recognition.

a. Noise Reduction

A median filter with a 5x5 window size will be used to reduce noisy pixels and enhance contrast between background and foreground pixels. On the word photo, a statistical-based smoothing method will be used to decrease noise and normalise the outer contour of the word by deleting

spurious pixels that may have been generated during the digitization process, creating irregularities in the word's outer contour.

b. Normalization

In automated cursive handwritten word recognition, normalizing is an important step. The goal of normalizing methods is to eliminate writing variations and generate uniform data. Skew normalisation and stroke thickness normalisation are two normalisation techniques used on Arabic text.

Baseline extraction and skew normalization

Some pages may not be directly fed into the scanner during the image capturing process, resulting in skewing of bitmapped pictures of these pages. This impact is seen as a slope in the text line relative to the x-axis. The skew detection/correction technique has proven in the literature to improve segmentation accuracy significantly. Baseline is important in Arabic text recognition since it helps us to differentiate between the main body of the word and dots and other diacritics. As there are numerous characters with the same form but differ only in the location and number of dots, the position of the dot (s) with relation to the baseline is crucial in the recognition of many Arabic characters. As a result, several researchers have focused on baseline extraction.

Stroke thickness Normalization

The thickness of a word in the input picture is adjusted in two steps. The entered word is first thinned. The binary image is then dilated with 3-pixels around the original thinned pixel, resulting in a consistent thickness of 3 pixels for every image, regardless of its initial thickness on the word image, to ensure appropriate contour generation by making the image stroke width at least three pixels broad.

c. Extraction of connected components

Many research projects start with the extraction of related aspects, which is motivated by the nature of Arabic writing, which consists of a series of elements (called pieces of Arabic word or sub-words). The purpose of word/sub-word extraction is to extract portions of Arabic words (PAWs), each of which is made up of one basic component (the main body) and any related dots/diacritics. Words/sub-words are extracted from text lines by extracting all of the line's associated components. The expectation-maximization procedure is then used to estimate a preliminary baseline. The key components of the PAWs are found to use this baseline. Secondary components include dots and other diacritics. A major component is assigned to each of the secondary components. Slant correcting is then implemented to the recovered word/sub-words.

(ii) Representation

The ensuing recognition steps are heavily affected by the choice of an adequate pattern representation. Context and skeleton representations are commonly utilized in addition to pixel representation. Contour representation converts each sub-word into a closed curve, whereas skeleton representation utilizes thinning to render the resulting image components one pixel thick.

Some methods use the image to create a skeleton or a list of contours. There are ways, however, that pass the entire pixel array representing the character/word picture to the next level of analysis. The improved binary text picture is thinned to decrease the quantity of data to be processed to the bare minimum required for running the segmentation algorithm and to make the process of extracting the essential feature points easier. The method of thinning adopted was based on the thinning procedure. Because it tolerates complexity and retains diacritic points, which are essential primitives for word discrimination, the thinning algorithm is used.

(iii) Segmentation

When opposed to non-connected writing styles like printed Latin, the linked character of Arabic lettering renders segmentation more complex. For the Arabic recognition system, the segmentation module is the most difficult stage. At various levels, researchers attempted to partition Arabic text. To segment Arabic text lines into words, a number of methods are described. Due to the overlapping structure of Arabic text, which contains both word gaps and interword breaks, the success rate was relatively low. Attempts to divide Arabic text into sub-words have also been attempted. Analysis of vertical projection, contour analysis, regularities and singularities, and topological characteristics have been used in methods for segmenting handwritten Arabic text into symbols.

(iv) Feature Extraction

The objective of feature is to quantify the patterns' qualities that are most related to a particular classification job. The selection of features to be extracted is an important step in ensuring the success of a recognition procedure, given subsequent processing will no longer change the actual picture, but rather the results supplied by the feature extraction module.

Loops, branch points, endpoints, and dots are structural qualities of writing that are intuitive. They're many in using a text image skeleton. In the case of handwritten Arabic recognition, structural characteristics provide a natural way of explicitly collecting dot information, which is needed to differentiate numerous letters. Statistical features are numerical measures calculated across pictures or image areas. Pixel densities, chain code direction histograms, moments, and Fourier descriptors are among them.

(v) Recognition

Character recognition's ultimate goal is to determine the class codes (labels) of character patterns. The goal of recognition becomes assigning each character pattern or word to a class from a predefined class set after segmenting character patterns or words from document images.

Segmentation-based (Analytical) and segmentation-free (Holistic) approaches to Arabic handwriting recognition have indeed been presented in the literature. The word is divided into smaller units (characters, graphemes, or primitives) in segmentation-based recognition systems, and then these units are recognised. Segmentation-free systems, on the other hand, treat the entire image as the unit of recognition. The systems proposed in the literature dealt with handwritten Arabic recognition at various levels, including character, numeral, sub-word, and word.

V CONCLUSION

This research mainly focussed on building unique techniques for optical character recognition (OCR) of handwritten text, with a particular focus on the Arabic language. The K-Nearest Neighbor (KNN), Support Vector Machines (SVM), and advanced bio inspired by neural networks approaches employed in this study were all examined. A thorough review of the literature on OCR with a focus on Arabic text has been published. As a result, a new research goal was formed in order to improve the efficiency of analysing Arabic handwritten character recognition.

REFERENCES

[1] Ahmad I, Fink GA (2016) Class-based contextual modeling for handwritten Arabic text recognition. In: 2016 frontiers in handwriting recognition (ICFHR), pp 554–559

[2] Baldominos A, Sa´ez Y, Isasi P (2019) A survey of handwritten character recognition with mnist and emnist. Appl Sci 2019:3169

[3] Ramzan M, Khan HU, Awan SM, Akhtar W, Ilyas M, Mahmood A, Zamir A (2018) A survey on using neural

network based algorithms for hand written digit recognition. Int J Adv Comput Sci Appl. https://doi.org/10.14569/ IJACSA.2018.090965

[4] El Qacimy, B.; Kerroum, M.A.; Hammouch, A. Word based Arabic handwritten recognition using SVM classifier with a reject option. In Proceedings of the 2015 15th international conference on intelligent systems design and applications (ISDA), Marrakech, Morocco, 14–16 December 2015; pp. 64–68.

[5] Asebriy, Z.; Raghay, S.; Bencharef, O.; Chihab, Y. Comparative systems of handwriting Arabic character recognition. In Proceedings of the 2014 Second World Conference on Complex Systems (WCCS), Agadir, Morocco, 10–12 November 2014; pp. 90–93.

[6] El Qacimy, B.; Hammouch, A.; Kerroum, M.A. A review of feature extraction techniques for handwritten Arabic text recognition. In Proceedings of the 2015 International Conference on Electrical and Information Technologies (ICEIT), Marrakech, Morocco, 25–27 March 2015; pp. 241–245.

[7] Patel, S.R.; Jha, J. Notice of Removal: Handwritten character recognition using machine learning approach-A survey. In Proceedings of the 2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), Visakhapatnam, India, 24–25 January 2015; pp. 1–5.

[8] Hussain, R.; Raza, A.; Siddiqi, I.; Khurshid, K.; Djeddi, C. A comprehensive survey of handwritten document benchmarks: Structure, usage and evaluation. EURASIP J. Image Video Process. 2015, 2015, 46.

[9] AlKhateeb, J.H. A database for Arabic handwritten character recognition. Procedia Comput. Sci. 2015, 65, 556–561.

[10] Khorsheed, M.S. Off-line Arabic character recognition–a review. Pattern Anal. Appl. 2002, 5, 31–45.

[11] l Moubtahij, H.; Halli, A.; Satori, K. Review of feature extraction techniques for offline handwriting arabic text recognition. Int. J. Adv. Eng. Technol. 2014, 7, 50.

[12] El-Sawy, A.; Hazem, E.B.; Loey, M. CNN for handwritten arabic digits recognition based on LeNet-5. In International Conference on Advanced Intelligent Systems and Informatics; Springer: Berlin/Heidelberg, Germany, 2016; pp. 566–575.

[13] Loey, M.; El-Sawy, A.; El-Bakry, H. Deep learning autoencoder approach for handwritten arabic digits recognition. arXiv 2017, arXiv:1706.06720.

[14] ElAdel, A.; Ejbali, R.; Zaied, M.; Amar, C.B. Dyadic multi-resolution analysis-based deep learning for Arabic handwritten character classification. In Proceedings of the 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), Vietri sul Mare, Italy, 9–11 November 2015; pp. 807–812.

[15] Myers, L. Aliterate Community College Remedial Students and Their Attitudes toward Reading: A Phenomenological Examination. Master's Thesis, California State University, Long Beach, CA, USA, 2013.

[16] El Qacimy, B.; Hammouch, A.; Kerroum, M.A. A review of feature extraction techniques for handwritten Arabic text recognition. In Proceedings of the 2015

International Conference on Electrical and Information Technologies (ICEIT), Marrakech, Morocco, 25–27 March 2015; pp. 241–245.

[17] Manisha, C.N.; Reddy, E.S.; Krishna, Y. Role of offline handwritten character recognition system in various applications. Int. J. Comput. Appl. 2016, 135, 30–33

[18]I. Palatnik de Sousa, ‘‘Convolutional ensembles for arabic handwritten character and digit recognition,’’ PeerJ Comput. Sci., vol. 4, p. e167, Oct. 2018, doi: 10.7717/peerj-cs.167.

[19]Azuraliza, A.B., Siti Rohaidah, A., Nurhafizah Moziyana, M.Y., Yaakub, M.R., 2017. Statistical analysis for validating ACO-KNN algorithm as feature selection in sentiment analysis. International Conference on Electronics and Communication System.

[20]Mudhsh MA, Almodfer R (2017) Arabic handwritten alphanumeric character recognition using very deep neural network. Information 8(3)

[21]Younis K (2018) Arabic handwritten characters recognition based on deep convolutional neural networks. Jordan J Comput Inform Technol (JJCIT)

[22] Peng, C., Limc, S., Chin Neoh, L., Zhang, K., Mistry, K., 2018. Feature selection using firefly optimization for classification and regression models. Decis. Support Syst. 106, 64–85.