Journal of Informatics and Mathematical Sciences
Vol. 8, No. 5, pp. 347–360, 2016
ISSN 0975-5748 (online); 0974-875X (print)
Published by RGN Publications

http://www.rgnpublications.com

Survey article

Existing Techniques in Arabic Characters Recognition (ACR)

Ayed Ahmad Radaideh* and Mohd Shafry Mohd Rahim

Department of Computer Science, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
***Corresponding author:** aidradaideh@yahoo.com

Abstract. Text or handwritten document writing recognition or understanding through characters whether online or off-line play an important role in many applications, e.g. recognizing text on bank cheques, editing old documents or translating online writing. The Arabic language characters recognition in Arabic text (hand written and printed text) is a challenging job, which is addressed by researchers in two different domains, i.e. online and off-line. The challenges are due to the cursive nature of the Arabic language. The approaches designed and adopted to achieve the target of Arabic characters recognition are also broadly classified into two, i.e. with segmentation and without segmentation. Machines can easily classify the characters if they are properly presented to it. The aim of this paper is to review critically state-of-the-art and to highlight the drawbacks in recent Arabic character recognition (ACR) for realization of the needs for robust online or off-line character recognition. This paper will properly enlighten the way for new researchers to improve or overcome the problems faced by previous studies and in addition provide a brief summary of ACR databases.

Keywords. Characters recognition; Characters segmentation; Features extraction; Classification; Arabic characters recognition

MSC. 68M12 (Network protocols)

Received: August 24, 2016 **Accepted:** November 9, 2016

1. Introduction

Every language is composed or formed by characters which are used to form words and words in turn form sentences. For the understanding of any language, whether the understanding is related to human (natural) or machines (artificial) both need the understanding or recognition

of the characters of that particular language. The process by which characters are recognized according to the predefined character class through machine (computer) is called character recognition. Due to the increase involvement of machines in daily life, it is also desired by the human community that computers are able to read the text regardless that it is handwritten (either written on a paper or on the screen of a computer) or in the form of printed document to enable the user to communicate with it. A great amount of research has been carried out for Chinese, Latin, and Arabic characters recognition during the last decades.

The Arabic language belongs to the semantic language family and among one of the mostly spoken languages in the world. Arabic letters recognition had not received enough care by researchers until 1975 when Nazif develop a system to recognize the Arabic language [50]. Arabic or Arabian language is a universal language, and it is the official spoken language in 25 different countries, with population more than 300,000,000 [12, 34, 39]. In addition, many Arabic characters not only used in the native Arabic language but also in another languages including Urdu, Farisi, Kardi, Jawi and Bahasa Melayu [61].

The Arabic language is semi-cursive because some of the basic characters are cursive while some are non-cursive, i.e. the total basic characters are 28, out of which 22 are cursive while the remaining 6 are non-cursive [51]. The demand for Arabic character segmentation, recognition and classification increased in 1980 and up to date the work is going for more and more improvement. In pattern recognition area, character recognition field is considered as one of the major active fields. However, the development of Arabic character recognition has not been focused from researchers as other languages, Chinese, Japanese and Latin recognition systems [61]. Generally, languages recognition is considered one of the most challenging problems in pattern recognition field [57].

There are many challenges faced by Arabic character recognition. As over and under segmentation in case of handwritten documents because the shape of the characters varies from writer to writer. The shape of letter has different form according to its position in the word making connecting point's variation in segmentation phase and more challengeable to recognize in the classifier. The cursive nature of Arabic letters yielding many overlapping letters. Spaces are existed that separate between words and strokes. These spaces are not uniform and varied in sizes making ambiguities to distinguish between stroke ends or word end in segmentation phase. Besides these, too many other challenges are faced by the ACR.

2. Arabic Characters Datasets

Many datasets encompassing images of samples of texts have been deployed for recognition handwriting issues. Nevertheless, the researchers still ask for more of them to evaluate and test their systems and making valuable comparison in their used techniques. The Technical University Braunschweig develops the IFN/ENIT-database in Germany. Another, database ENIT is prepared in Tunisia. The first version of this database contains 26,459 images of the 937 names of cities and towns in Tunisia. The writers were 411. The database includes 115585 pieces of Arabic words (PAWs) and 212211 characters. The images are grouped into four groups. Many researchers of Arabic handwritten text recognition have used this database. A competition of Arabic handwritten text recognition was conducted using this database in

2005, 2007 and 2009. However, the database contains only the names of cities and thus has vocabulary limitations. Adnan and Amin developed another database consisting of 4800 isolated Arabic handwritten characters by Amin [14].

Al-Ohali et al. [51] produced a database for Arabic handwritten cheques. Legal and courtesy amounts were extracted from 3000 cheques of "Al-Rajhi Banking and Investment Corp" of Saudi Arabia. The database contains 2499 legal amounts, 501 courtesy amounts written in Indian/Arabic digits. The Arabic sub-words used are 29498 in legal amounts and 15175 Indian/Arabic digits. The extracted information was from previously collected checks and it is the most natural database for Arabic bank cheques analysis and recognition. The mentioned database is used for Arabic handwritten digit recognition but has limited vocabulary regarding word recognition.

Mozaffari et al. [46] developed a database of handwritten Farsi/Arabic numbers and characters called IFHCDB. The database contains grayscale images of 52380 characters and 17740 numerals. The images were scanned at 300 dpi from the Iranian high school and guidance school entrance exam forms during 2004- 2006. This database lacks natural Arabic/Farsi handwritten text. In addition, the distribution of different letters and characters is not uniform.

Khosravi and Kabir presented a handwritten Farsi digit database in [35]. It contains 102352 digits extracted from about 12000 registration forms filled by B.Sc. and senior high school students. In 2007, Applied Media Analysis released a database called Arabic-Handwritten-1.0. This database includes 5000 images, in which there are 200 documents of Arabic manuscripts, notes, diagrams, poems, forms and Indian and Arabic numerals written by 25 writers. This database is a commercial database.

El-Sherif and Abdelazeem [25] developed a database of Arabic numerals called ADBase. ADBase is composed of 70,000 digits written by 700 writers. Each writer wrote each digit (0-9) ten times. The database is partitioned into two sets: a training set consisting of 60,000 digits samples and a test set of 10,000 digits samples. This database is the most comprehensive database for Arabic handwritten digits.

Alamri et al. [8] produces a database for Arabic handwriting recognition. The database contains 284 samples of Arabic dates written by 284 writers. It also contains 46800 isolated digits, 13439 numerical strings, 21426 isolated letters, 11375 samples of 70 Arabic words and 1640 samples of special symbols dataset written by 328 writers.

Mahmoud [47] developed a database of Arabic (Indian) digits written by 44 writers. Each writer wrote 48 samples of each digit resulting in a database of 21,120 samples. This database is limited in the number of writers and does not include the natural variations of Arabic handwritten digits.

Plamondon and Srihari [53] presented a database consisting of approximately 20,000 words (10 writers have written 10 pages of text, each includes between 150 and 200 words). The database is limited and not freely available.

Kherallah et al. [36] presents AHD/AUST, a database of Arabic handwritten words. The database contains 12300 Arabic words written by 82 different writers. They produces a database called the On/Off (LMCA) dual Arabic database. This database is used for both on-line and off-line recognition and consists of letters, words and digits. The database has 30000 digits,

100000 letters and 500 words by 55 writers. For each character and word, handwritten trajectory is collected by recording the (x, y) co-ordinates for on-line recognition. Off-line procedure is based on the collection of images of the handwritten trajectory.

Motivated from the IfN/ENIT database, Mozaffari et al. [49] produced a database called IFN/Farsi. This database consists of 7271 binary images of 1080 Iranian province/city names written by 600 writers. This database has the same limitations as IFN/ENIT.

Ziaratban et al. [65] presented a Farsi Handwritten Text database (FHT). This database was created by filling 1000 forms by 250 writers. An average, each form contains 6.45 text lines. In total, the database contains 106600 handwritten words. This is one of 40th recent attempt in collecting large handwritten Arabic/Farsi text database. Although FHT is a Farsi text database, it also utilized by researchers in Arabic text recognition.

Al-Muhtaseb et al. [13] presented an algorithm to produce a minimal Arabic text that covers the different basic shapes of Arabic alphabet. They used several lexicons and Arabic text to produce a minimal Arabic text, which can be used for collecting Arabic handwritten text database. However, the minimal printed text produced by their algorithm is not natural and thus may not reflect the natural way of writing by a person. Next, we presented the tabular summary of the notable databases as discussed above with the details of the language and number of writers in Table 1 available database for ACR.

Table 1. Presents commercial database for Arabic handwritten text recognition

Dataset	**Explanation**	**#writers**
IFN/ENIT	26,459 images of Tunisian cities names	411
Al-Isra	37,000 words, 2500 signatures, 10,000 digits, 500 sentences	500
Applied Media Analysis	5000 images, 200 documents	-
AHDB	10,000 words, handwriting pages	100
Khedher et al.	48 pages of texts	48
Alamri et al.	46,800 digits, 21,426 letters, 13,439 numerical strings, 11,375 words, 1640 special symbols	328
IFN/Farsi	7,271 images of 1,080 Iranian province/city names	600
CENPARMI	3,000 checks	-
FHT	106,600 words	250
Mahmoud	21,120 digits	44
CEDAR	20,000 word	10
IFHCDB	17,740 numerals and 52,380 characters	-
ON/OFF LMCA	30.000 digits, 100000 letters and 500 word	55
ADBase/MADBase	70,000 digit	700
Khosravi et al.	102,352 Farisi digits	12000 forms

3. Arabic Characters Recognition (ACR) Approaches

The ACR adopted and developed techniques are categorized according to its domains and approaches. The domains include on line and off-line while in both the domains process is accomplished with and without segmentation approach. The proper categorization is presented in Figure 1 while in Figure 2 the general methodology for accomplishing the process of ACR.

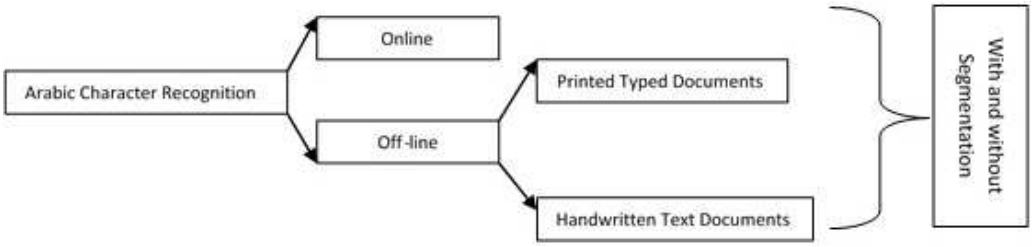

Figure 1. Domains and Approaches of ACR

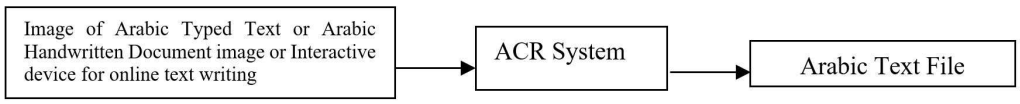

Figure 2. Arabic Character Recognition system general overview

3.1 Pre-processing

The Improvement in the preprocessing stage has significant impact on the efficiency and reliabilityof the segmentation, feature extraction and classification processes [15, 26]. More commonly, the OCR preprocessing stage composed of binarization, filtering, slant detection and correction, smoothing, interpolation, thinning (Skelotonization) and base line detection. Not all these listed operations are compulsory, but it depends on the working of OCR system.

Figure 3 shows the operations for the ACR preprocessing. It starts in off-line documents or manuscripts analysis for ACR from binarization, followed by skeltonization, further slant detection, and baseline detection. However, these are not hard fast rules, but it varies from study to study according to their need. Some previous studies in the same domain are without segmentation, they extract features and then did classification, these studies have no need to detect slants and base lines.

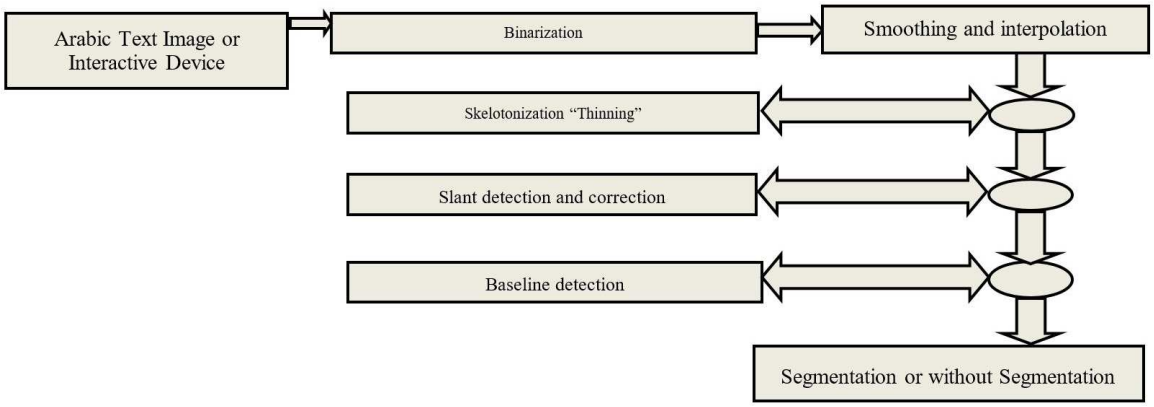

Figure 3. Some frequently used pre-processing operations

3.2 Online and Off-Line ACRs with Segmentation Approach

It is considered one of the most major problems in text recognition phases to segment a word to characters [31]. Some systems do not use segmentation phase to segment the words of text image into their graphemes or isolated recognizable characters [62]. However, approaches used in segmentation suffer from many difficulties due to the cursive nature of Arabic. These difficulties include: Misplaced segmentation in character overlapping, over segmentation (segmenting one character into sub-parts), under segmentation as in the ligatures where two letters are connected to each other in a way that they cannot be separated horizontally [54].

Segmentation in text images is considered as important and challenging issue in OCR system especially in Arabic letters. It has direct significant effects on the following stages; feature extraction and classification process [57, 61]. Many segmentation techniques have been developed. However, Arabic text with its complex nature remain as an obstacle the researchers are facing [61].

In Arabic text, characters recognition approaches are classified generally into two distinct approaches: segmentation-free approach (Holistic Approach) and segmentation-based approach. In the former, the entire word is recognized as a set of characters without prior segmentation in the original word [23, 32]. In the same approach, the Arabic text is just segmented to words without segmenting into the word's graphemes [61]. In the later approach, a word or sub-word is segmented into tokens. These tokens might be grapheme, set of graphemes and strokes [61]. The segmentation-based stage involves several steps: segmentation of lines, segmentation of words/sub-words and segmentation to graphemes. This stage is considered as an important stage because it has a direct impact on the identification rate of the whole recognition system [5, 20].

3.2.1 Segmentation into Lines

Normally, the text image after scanning may have one or more paragraph, which consists of several lines. These lines should be extracted by separating them from each other without losing the text contents. There are many techniques used to accomplish this task, for instance, horizontal projection, vertical projection [7, 53]

3.2.2 Segmentation into Words/sub-words

After segmenting the lines, the next important step is to segment the words and sub-words from the lines separately. The gaps between the words/sub-words play an important role to carry out this task. Furthermore, some Arabic characters cannot be joined with the subsequent part in the same word as in the word عادل which yields to produce sub-words after applying this segmentation task. The existence of longer spaces separating between words than spaces between sub-words can be helpful to distinguish between them. Hence, it is noted generally in the handwriting of many writers that they leave more space among the words as compare to characters [37].The distances among the words(connected components) is far better for segmentation of words [11].

3.2.3 Segmentation into characters/graphemes

This type of segmentation aims to segment words/sub-words into single characters or graphemes. As described in [38], the critical points are determined between characters, i.e. the end points and start points of each character are identified. However, Arabic character segmentation still have many challenging issues, i.e. Arabic handwriting is considered as cursive nature makes segmentation process of individual characters difficult [61] and [5] described various techniques to be used for segmenting a word into characters:

- *Segmentation based on Vertical Projection*:
  This technique considered the thickness of linking stroke and how the characters are connected in the word, this connection should be less than the other linking strokes. Furthermore, the technique the two-dimensional information in the text image will be reduced into one dimensional. Many previous studies proposed goodapproaches for vertical based segmentation [38, 52, 63] but still the process needs improvement in terms of accuracy.
- *Segmentation based on Contours*:
  This technique focuses on the contours to be traced for the word. It traces the outer part of the contour of a word. Sari et al. [58] suggested an approach for the same purpose via analyzing and morphological rules. First, the local minimum point in the lower part of contour is searched. Next, the morphological rules are used for taking decision about the point that it is the segmentation point or not? [45] proposed an algorithm that aims to detect the local maxima and minima in the lower and upper contour respectively. These detected points will mark then as intended boundaries for letters. A set of rules based on cursive scripts are applied to the boundaries points to and eliminate the unrelated boundary points. A comparison process between lower and upper boundaries points is performed to minimize the overlaps in characters.
- *Segmentation based on Skelotonization*:
  Thinning or skelotonization is a technique provides important information about the orientation and shape of the character. Many approaches were developed in the literature regarding how to extract the skeletons of the word? Al-Sadoun and Amin [16] proposed an approach for word tracing in the text image to extract the skeleton and then scan from right to left to find all critical points. Tellache et al. [59] applied skelotonization on the word in order to detect all edge points, break-points, and end-points of the word.
- *Segmentation Technique via Morphological Operators*:
  Closing operator followed by opening operator are adopted to segment the word into character, more than one character or part of a character. The approach is followed by [48] in implementation for segmentation of handwritten Arabic words. They further suppose that by applying the opening operator they will get the vertical strokes at the start and end of a character as singularities. Further, the input image is traced from right to left to identify the regularities having required information to connect the subsequent characters, i.e. to identify the segmentation points. After obtaining the regularities, they are subtracted from the singularities of the input image to disconnect a character from

a subsequent character. The problem of overlapping characters in Arabic text remains challenging issue in Arabic character recognition, which needs improvement [61].

- *Segmentation via ANNs*:
  Artificial Neural Networks are trained to classify the actual points suitable for segmentation. Hamid and Haraty [30], developed a hybrid approach for segmentation of characters from Arabic handwritten documents. Further, they divide their approach in two main components. In the first component, they use a heuristic method to scan the handwritten images of text to extract block of characters (BCs). For this purpose, they utilize topographic features of subsequent block of characters to identify pre-segmentation points. In the second component, they use ANN, to verify the extract validity [61].

3.3 Online and Off-Line ACRs Features Extraction Approaches (without segmentation)

Information that are obtained from the text image or word or even from a character called features. Features can also be defined as any information come from pixels intensity, mathematical properties in the shape of lines or curves and shape of data. The mentioned information are significant to be extracted and passed to the recognizer for building models and classification. Feature selection and extraction plays a fundamental role in pattern recognition [22]. Therefore, selecting the appropriate features having information are extracted from text images, and this is important job in Arabic character recognition. However, selecting the important features among others is a challenging problem for the purpose to exclude irrelevant and redundant features to use.

Typically, this selection can also decrease the complexity of the system which makes processing time and recognition rate more efficient and accurate [28]. Features are also important in the AOCR system. They have great impact on the classification stage [39]. This phase is used to analyze an Arabic text after being segmentation of characters or directly in holistic approaches, the segmented features or the combination of many segmented features of text image will be used for the classification purposes so that to improve the overall recognition rate [34, 66].

There are many methods used for Arabic text feature extraction which are categorized generally into three main categories: Structural features, Statistical features, Global Transformations. The structural features are based on text image physical appearance including the loops, zigzags, intersection points, width, height, dots, number of crossing points, curves, lines, junctions, word length, position and ascenders sequence described in [2, 17, 27, 28, 33, 43, 66]. Statistical features are based on statistical methods and are considered the fastest and efficient methods for feature extraction but are sensitive to noise.

The statistical methods used in Arabic text image for feature extraction include the following: (1) Zoning, (2) Character loci, (3) Moments and Crossings [9, 19, 56]. The Global Transformation methods work on the concept to shorten the text in order to obtain good features. The methods used under this category for Arabic text features extraction include Vertical and Horizontal projections, Hough transform, Gabor transform and coding [55, 60].

3.3.1 Feature extraction of Characters

Al-Emami and Usher [10] described the different classes of four direction of the slope linking between starting point and the points used in segmentation of Arabic character. Figure shows these directions and slope. Continuous segments having the same direction class will be merged all together to form one unit class.

Halavati et al. [29] defined a fuzzy feature set based on the ratio length of segment from pattern of handwriting, and different shapes features are extracted from these segments. The features are arc, lines, or loops.

3.4 Classification methods in On-line and Off-line ACRs

In character recognition systems, most of the researchers use ANN (*Artificial Neural Networks*) in their research. ANN provides great fault tolerance than any another conventional classifiers; because it have large number of connections in the network. In addition, its performance does not impair in the network even when some of connections or/and nodes are damaged [56]. In classification, it assigns some unknown feature to a predefined class for classification purpose.

Arabic recognition systems can identify the text in two approaches; either Global (Holistic) or Analytic approach. The formal strategy can identify the whole words or sub words, and it does not involve segmentation as well, besides it works only on limited number of vocabularies [6, 24, 57]. The analytic strategy identifies the segmented features, as well it requires segmentation, and can be applied on unlimited vocabularies [44, 49]. Many classification methods have been used in recognizing Arabic text as statistical techniques, template matching, neural networks, syntactic techniques and Hidden Markov Model [18, 24, 57].

3.4.1 Classification Approaches

Support Vector Machine (SVM) with Radial Basis Function (RBF) as the kernel was used in [4] for Arabic numeral recognition. Alaei et al. [3] experimented with SVM with linear, Gaussian and polynomial kernels fore numeral recognition. The authors recounted that SVM with Gaussian kernel can give the best combination. SVM is used for Arabic numeral recognition. Angle, distance, horizontal, and vertical span features are used in [42]. Two stages with extensive parameter estimation technique was used to predicate the best values for the (SVM) parameters. A dataset of (44) writers with (48) samples for each digit, with total of (21120) samples was used in (SVM) parameter prediction. Therefore, the dataset splitto (4) sub-sets: (3) is used for the validation training and in turn, and the 4thused for the testing purpose. A recognition accuracy achieved is 99.39% and is reported in [42].

In the other hand, neural networks were used for recognizing Arabic word in [6]. Ziaratban et al. [64] utilized neural networks for numeral recognition. Firstly, they found out the average of all images of all numerals then. Secondly, they extract (20) templates from those computed average images. For extracting features, the templates are used for matching the numeral images to find positions of the best matching. After that, they coordinate the best match and the amount of matching used for features. Besides, Learning Vector Quantization (LVQ) neural network was used in (Ali 2008) for recognizing Arabic handwritten characters.

The two-tier approach is also used, and it is presented in [1], in this work, they takes the (PAWs) from Arabic text words as different alphabets. They divide the word's recognition

problem to two concurrent problems: (1) to find the best possible mapping words from PAWs to words. (2) To find the best possible mapping from characters in (PAWs). Two NN-based classifiers used in [1] to identify the (PAWs) in a word. Then beam-search used for finding the best words which are matching. Results were stated on IFN/ENIT dataset with 11.06% word error rate on set-d. Cheikh et al. [21] adopted the use of linguistic characteristics in Arabic language for recognizing Arabic words. Their work is starting in partitioning the Arabic word, in each word it should be partitioned to a prefix, suffix and a radical. Where is the radical to be a derivation from a word root according to some scheme.

Cheikh et al. [21] used two transparent neural networks (TNN). In the first TNN network is used to take a word then extract its root. While in the second TNN network was trained in a way that to ignore the letter's root, then to extract the scheme in the radical of the word's root. For a limited database of (2250) words in machine printed, the stated recognition rate for the two networks were 91% and 76.5% for top-4 choices respectively. Multidimensional recurrent neural networks (MDRN) and connectionist temporal classification for handwritten word recognition utilized in Graveset al. in (2009).In their work a globally trained off-line handwriting recognizer is introduced, which consider the raw pixel data as an input to the network. They found 91.4% accuracy rate for set-f of IFN/ENIT database. Further, some latest literature in both the domains is summarized along with the results description in Table 1.

4. Conclusion

In this paper, we are intended to highlights the gaps or limitation left by the previous researcher and to concentrate more that today still more space is available in terms of accuracy and efficiency in Arabic Character Recognition (ACR) systems. The limitations in both the domains are highlighted, and the way is enlightened for the fresh researchers to guide them towards the right path. The databases are also discussed in detail for the selection of a proper database selection for the future.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] A. AbdulKader, A two-tier Arabic offline handwriting recognition based on conditional joining rules, in *Arabic and Chinese Handwriting Recognition*, pp. 70 – 81, Springer, 2008.

[2] I.S.I. Abuhaiba, A discrete Arabic script for better automatic document understanding, (2003).

[3] A. Alaei, U. Pal and P. Nagabhushan, Using modified contour features and SVM based classifier for the recognition of Persian/Arabic handwritten numerals, in *IEEE Seventh International Conference on Advances in Pattern Recognition (ICAPR'09)* (2009), pp. 391 – 394.

[4] H. Alamri, C.L. He and C.Y. Suen, A new approach for segmentation and recognition of Arabic handwritten touching numeral pairs, in *International Conference on Computer Analysis of Images and Patterns* (2009), pp. 165 – 172, Springer.

[5] Y.M. Alginahi, A survey on Arabic character segmentation, *International Journal on Document Analysis and Recognition (IJDAR)* Vol. 16 (2) (2013), 105 – 126.

[6] S. Alma'adeed, C. Higgins and D. Elliman, Off-line recognition of handwritten Arabic words using multiple hidden markov models, *Knowledge-Based Systems* **17** (2) (2004), 75 – 79.

[7] Z. Al Aghbari and S. Brook, Hah manuscripts: A holistic paradigm for classifying and retrieving historical Arabic handwritten documents, *Expert Systems with Applications* **36** (8) (2009), 10942 – 10951.

[8] H. Alamri, J. Sadri, C.Y. Suen and N. Nobile, A novel comprehensive database for Arabic off-line handwriting recognition, in *Proceedings of 11th International Conference on Frontiers in Handwriting Recognition* (ICFHR) (2008), Vol.**8**, pp. 664 – 669.

[9] B. Al-Badr and S.A. Mahmoud, Survey and bibliography of Arabic optical text recognition, *Signal Processing* **41** (1) (1995), 49 – 77.

[10] S. Al-Emami and M. Usher, On-line recognition of handwritten Arabic characters, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12** (7) (1990), 704 – 710.

[11] J.H. AlKhateeb, J. Jiang, J. Ren and S. Ipson, Component-based segmentation of words from handwritten Arabic text, *International Journal of Computer Systems Science and Engineering*, **5** (1) (2009).

[12] J.H. AlKhateeb, J. Ren, S.S. Ipson and J. Jiang, Knowledge-based baseline detection and optimal thresholding for words segmentation in efficient pre-processing of handwritten Arabic text, in *IEEE Fifth International Conference on Information Technology: New Generations, 2008 (ITNG2008)*, pp. 1158 – 1159.

[13] H.A. Al-Muhtaseb, S.A. Mahmoud, R.S. Qahwaji, M. Demiralp, N.A. Baykara and N.E. Mastorakis, A novel minimal arabic script for preparing databases and benchmarks for arabic text recognition research, in *WSEAS International Conference, Proceedings: Mathematics and Computers in Science and Engineering*, No. 8, World Scientific and Engineering Academy and Society, 2009.

[14] A. Amin, Recognition of hand-printed characters based on structural description and inductive logic programming, *Pattern Recognition Letters* **24** (16) (2003), 3187 – 3196.

[15] H. Al-Rashaideh, Preprocessing phase for Arabic word handwritten recognition, *Information Process (Russian)* **6** (1) (2006).

[16] H.B. Al-Sadoun and A. Amin, A new structural technique for recognizing printed Arabic text, *International Journal of Pattern Recognition and Artificial Intelligence* **9** (01) (1995), 101 – 125.

[17] A.T. Al-Taani, An efficient feature extraction algorithm for the recognition of handwritten Arabic digits, *International Journal of Computational Intelligence* **2** (2) (2005), 107 – 111.

[18] A. Amin and J.F. Mari, Machine recognition and correction of printed Arabic text, *IEEE Transactions on systems, man, and cybernetics* **19** (5) (1989), 1300 – 1306.

[19] I. Bazzi, R. Schwartz and J. Makhoul, An omnifont open-vocabulary OCR system for English and Arabic, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21** (6) (1999), 495 – 504.

[20] R.G. Casey and E. Lecolinet, Strategies in character segmentation: A survey, in *Proceedings of the IEEE Third International Conference on Document Analysis and Recognition*, Vol. 2, pp. 1028 – 1033, 1995.

[21] I.B. Cheikh, A. Belaïd and A. Kacem, A novel approach for the recognition of a wide Arabic handwritten word lexicon, in *IEEE 19th International Conference on Pattern Recognition (ICPR2008)*, pp. 1 – 4, 2008.

[22] N. Das, A.F. Mollah, R. Sarkar and S. Basu, A comparative study of different feature sets for recognition of handwritten Arabic numerals using a multi layer perceptron, *arXiv preprint arXiv:1003.1894*, 2010.

[23] M. Dehghan, K. Faez, M. Ahmadi and M. Shridhar, Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM, *Pattern Recognition* **34** (5) (2001), 1057 – 1065.

[24] A. Dehghani, F. Shabini and P. Nava, Off-line recognition of isolated Persian handwritten characters using multiple hidden Markov models, in *Proceedings of International Conference on Information Technology: Coding and Computing*, pp. 506 – 510, 2001.

[25] E.A. El-Sherif and S. Abdelazeem, A two-stage system for Arabic handwritten digit recognition tested on a new large database, in *Artificial Intelligence and Pattern Recognition*, pp. 237 – 242, 2007.

[26] F. Farooq, V. Govindaraju and M. Perrone, Pre-processing methods for handwritten Arabic documents, in *IEEE Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 267 – 271, 2005.

[27] H. Goraine, M. Usher and S. Al-Emami, Off-line Arabic character recognition, *Computer* **25** (7) (1992), 71 – 74.

[28] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *Journal of Machine Learning Research* **3** (March 2003), 1157 – 1182.

[29] R. Halavati, S.B. Shouraki and S. Hassanpour, Evolution of multiple states machines for recognition of online cursive handwriting, in *IEEE 2006 World Automation Congress*, pp. 1 – 6, 2006.

[30] A. Hamid and R. Haraty, A neuro-heuristic approach for segmenting handwritten Arabic text, in *ACS/IEEE International Conference on Computer Systems and Applications*, pp. 110 – 113, 2001.

[31] C.J. Hilditch, Comparison of thinning algorithms on a parallel processor, *Image and Vision Computing* **1** (3) (1983), 115 – 132.

[32] M.S. Khorsheed, Hmm-based system for recognizing words in historical Arabic manuscript, *International Journal of Robotics & Automation* **22** (4) (2007), p. 294.

[33] M.S. Khorsheed and W.F. Clocksin, Structural features of cursive Arabic script, in *BMVC*, pp. 1–10, Citeseer, 1999.

[34] M.S. Khorsheed and W.F. Clocksin, Segmentation-free word recognition for Arabic handwriting, in *The International Conference on Pattern Recognition ICPR*, 2000.

[35] H. Khosravi and E. Kabir, Introducing a very large dataset of handwritten Farsi digits and a study on their varieties, *Pattern Recognition Letters* **28** (10) (2007), 1133 – 1141.

[36] M. Kherallah, A. Elbaati, H.E. Abed and A.M. Alimi, The on/off (lmca) dual Arabic handwriting database, in *11th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2008.

[37] G. Kim, V. Govindaraju and S.N. Srihari, Architecture for handwritten text recognition systems, *Series in Machine Perception and Artificial Intelligence* **34** (2000), 163 – 172.

[38] L. Lorigo and V. Govindaraju, Segmentation and pre-recognition of Arabic handwriting, in *IEEE Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 605 – 609, 2005.

[39] L.M. Lorigo and V. Govindaraju, Offline Arabic handwriting recognition: a survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28** (5) (2006), 712 – 724.

[47] S. Mahmoud, Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models, *Signal Processing*, **88** (4) (2008), 844–857.

[49] S. Mozaffari, K. Faez and M. Ziaratban, Structural decomposition and statistical description of Farsi/Arabic handwritten numeric characters, in *IEEE Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 237 – 241, 2005.

[42] S.A. Mahmoud and S.O. Olatunji, Automatic recognition of off-line handwritten Arabic (Indian) numerals using support vector and extreme learning machines, *International Journal of Imaging* **2** (A09) (2009), 34 – 53.

[43] A. Malaviya and L. Peters, Extracting meaningful handwriting features with fuzzy aggregation method, in *Proceedings of the Third International Conference on Document Analysis and Recognition*, Vol. 2, pp. 841 – 844, 1995.

[44] M.G. Mostafa, An adaptive algorithm for the automatic segmentation of printed Arabic text, in *17th National Computer Conference*, pp. 437 – 444, Citeseer, 2004.

[45] K. Mostafa and A.M. Darwish, Robust baseline-independent algorithms for segmentation and reconstruction of arabic handwritten cursive script, in *Electronic Imaging'99*, pp. 73 – 83, International Society for Optics and Photonics, 1999.

[46] S. Mozaffari, H. El Abed, V. Märgner, K. Faez and A. Amirshahi, Ifn/Farsi-database: a database of Farsi handwritten city names, in *International Conference on Frontiers in Handwriting Recognition*, 2008.

[47] S. Mahmoud, Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models, *Signal Processing* **88** (4) (2008), 844 – 857.

[48] D. Motawa, A. Amin and R. Sabourin, Segmentation of Arabic cursive script, in *ICDAR*, Vol. 97, pp. 625 – 628, 1997.

[49] S. Mozaffari, K. Faez and M. Ziaratban, Structural decomposition and statistical description of Farsi/Arabic handwritten numeric characters, in *IEEE Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 237 – 241, 2005.

[50] A Nazif, *A System for the Recognition of the Printed Arabic Characters*, Master's Thesis (2nd Edition), Faculty of Engineering, Cairo University (1975).

[51] Y. Al-Ohali, M. Cheriet and C. Suen, Databases for recognition of handwritten Arabic cheques, *Pattern Recognition* **36** (1) (2003), 111 – 121.

[52] B. Parhami and M. Taraghi, Automatic recognition of printed Farsi texts, *Pattern Recognition* **14** (1-6) (1981), 395 – 403.

[53] R. Plamondon and S.N. Srihari, Online and off-line handwriting recognition: a comprehensive survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** (1) (2000), 63 – 84.

[54] M.Y. Potrus, U.K. Ngah and B.S. Ahmed, An evolutionary harmony search algorithm with dominant point detection for recognition-based segmentation of online Arabic text recognition, *Ain Shams Engineering Journal* **5** (4) (2014), 1129 – 1139.

[55] S. Saadallah and S. Yacu, Design of an arabic character reading machine, in *Proceedings of Computer Processing of Arabic Language, Kuwait*, pp. 123 – 135, 1985.

[56] H.Y.Y. Sanossianm An arabic character recognition system using neural network, in *Proceedings of the 1996 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing VI*, pp. 340 – 348, 1996.

[57] R. Safabakhsh and P. Adibi, Nastaaligh handwritten word recognition using a continuous-density variable-duration hmm, *Arabian Journal for Science and Engineering* **30** (1) (2005), 95 – 120.

[58] T. Sari, L. Souici and M. Sellami, Off-line handwritten arabic character segmentation algorithm: Acsa, in *Proceedings of IEEE Eighth International Workshop on Frontiers in Handwriting Recognition*, pp. 452 – 457, 2002.

[59] M. Tellache, M. Sid-Ahmed and B. Abaza, Thinning algorithms for Arabic OCR, in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Vol. 1, pp. 248 – 251, 1993.

[60] F. Zaki, S. Elkonyaly, A. Elfattah and Y. Enab, A new technique for Arabic handwriting recognition, in *Proceedings of the 11th International Conference for Statistics and Computer Science, Cairo, Egypt*, pp. 171 – 180, 1986.

[61] A.M. Zeki, The segmentation problem in Arabic character recognition the state of the art, in *IEEE 2005 International Conference on Information and Communication Technologies*, pp. 11 – 26.

[62] L. Zheng, Machine printed Arabic character recognition using s-gcm. in *IEEE 18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 2, pp. 893 – 896, 2006.

[63] L. Zheng, A.H. Hassin and X. Tang, A new algorithm for machine printed Arabic character segmentation, *Pattern Recognition Letters* **25** (15) (2004), 1723 – 1729.

[64] M. Ziaratban, K. Faez and F. Faradji, Language-based feature extraction using template-matching in Farsi/Arabic handwritten numeral recognition, in *IEEE Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Vol.1, pp. 297 – 301, 2007.

[65] M. Ziaratban, K. Faez and F. Bagheri, Fht: An unconstraint Farsi handwritten text database, in *IEEE 10th International Conference on Document Analysis and Recognition*, pp. 281 – 285, 2009.

[66] A.B.C. Zidouri, S. Chinveeraphan and M. Sato, Structural features by MCR expression for printed Arabic character recognition, in *International Conference on Image Analysis and Processing*, pp. 557 – 562, Springer, 1995.