World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 1, No. 5, 184-192, 2011

Offline Arabic Text Recognition – An Overview

Atallah Mahmoud AL-Shatnawi
Department of Computer Science
Irbid National University
Irbid, Jordan
atallahshatnawi@gmail.com

Farah Hanna AL-Zawaideh
Department of Information System
Irbid National University
Irbid, Jordan
fomh2009@Gmail.com

Safwan AL-Salaimeh
Department of Computer Science
Irbid National University
Irbid, Jordan,
salsalaimeh@yahoo.com

Khairuddin Omar
Center for Artificial Intelligence Technology
Universiti Kebangsaan Malaysia
Malaysia
ko@ftsm.ukm.my

Abstract— This paper provides and discusses an overview about the offline **A**rabic **O**ptical **C**haracter **R**ecognition (AOCR) system. It also provides and discuses the challenges that must be considered in designing or choosing a cretin method for the AOCR system. Recognition of Arabic characters is more difficult than Latin or Chinese language .The typical AOCR system consists of five components: image acquisition, preprocessing, segmentation, feature extraction and classification (recognition). Each of those contributes to the final recognition rate to improve of the AOCR. In this paper, the challenges in recognition the Arabic written text are explained and discussed. As well as, the operations of offline AOCR system stages are discussed and studied in detail. The operations of AOCR preprocessing stage are also provided and discussed. The AOCR stages drawbacks and advantages are discussed in details.

Keywords- Arabic Text; Character recognition; Pre-processing; Segmentation; Feature Extraction; Classification.

I. INTRODUCTION

The goal of the Arabic Optical Character Recognition (AOCR) systems is to transform the input data (pattern of data), such as text written document on manuscript, text typed on document or online writing into a digital format. This can be manipulated by word processing software [61] [14]. In pattern recognition field, languages recognition is considered as one of the most complicated problem in Artificial Intelligent field [49]. Generally, Arabic Recognition can be done offline or online. In offline recognition, papers, manuscripts or documents are scanned or captured, and finally are manipulated by AOCR system. In online recognition application takes place during the writing process, many systems where developed for manipulating online AOCR such as [51] [2]. The online recognition system however is beyond the scope of this paper.

Arabic language is universal and it is a formal language for 25 countries, of population over than 300 million [31] [44] [36]. Additionally, many Arabic characters are used in different languages such as Ardu, Farsi, Jawi, Kardi [8].

This paper is organized as follows. Section 2 describes the Arabic written characteristics. Sections 3 clarify the offline AOCR system. Section 4 provides the discussion. Finally, section 5 presents the conclusion and the future direction.

II. ARABIC WRITTEN CHARACTERISTICS

It has been argued that recognition of Arabic characters is more difficult than others, such as Latin and Chinese [31] [61] [10]. The recognition difficulties can be referred to the following reasons:

1- Arabic language is written cursively, and it has 28 characters and each character is written between two to four shapes according to its location in the word (see table 1).

2- The Arabic language is written from right to left. Fig 1 shows the Arabic printed sentence “Alhoma Sali Ala

Sidina Mohammed" (اللهم صلي على سيدنا محمد) "is written from right to left using the Andalus font types.

TABLE 1. Shapes of Arabic characters in different positions

No	Character Name		Isolated	Connected		
				Beginning	*Middle*	*End*
1	**Alif**	**ألف**	ا	ا	ـا	ـا
2	**Baa**	**باء**	ب	بـ	ـبـ	ـب
3	**Taa**	**تاء**	ت	تـ	ـتـ	ـت
4	**Thaa**	**ثاء**	ث	ثـ	ـثـ	ـث
5	**Jeem**	**جيم**	ج	جـ	ـجـ	ـج
6	**Haa**	**حاء**	ح	حـ	ـحـ	ـح
7	**Khaa**	**خاء**	خ	خـ	ـخـ	ـخ
8	**Daal**	**دال**	د	د	ـد	ـد
9	**Thaal**	**ذال**	ذ	ذ	ـذ	ـذ
10	**Raa**	**راي**	ر	ر	ـر	ـر
11	**Zaay**	**زاي**	ز	ز	ـز	ـز
12	**Seen**	**سين**	س	سـ	ـسـ	ـس
13	**Sheen**	**شين**	ش	شـ	ـشـ	ـش
14	**Saad**	**صاد**	ص	صـ	ـصـ	ـص
15	**Dhaad**	**ضاد**	ض	ضـ	ـضـ	ـض
16	**Ttaa**	**طاء**	ط	طـ	ـطـ	ـط
17	**Dthaa**	**ظاء**	ظ	ظـ	ـظـ	ـظ
18	**Ain**	**عين**	ع	عـ	ـعـ	ـع
19	**Ghen**	**غين**	غ	غـ	ـغـ	ـغ
20	**Faa**	**فاء**	ف	فـ	ـفـ	ـف
21	**Qaf**	**قاف**	ق	قـ	ـقـ	ـق
22	**Kaf**	**كاف**	ك	كـ	ـكـ	ـك
23	**Lam**	**لام**	ل	لـ	ـلـ	ـل
24	**Mem**	**ميم**	م	مـ	ـمـ	ـم
25	**Noon**	**نون**	ن	نـ	ـنـ	ـن
26	**Haa**	**هاء**	ه	هـ	ـهـ	ـه
27	**Wow**	**واو**	و	و	ـو	ـو
28	**Yaa**	**ياء**	ي	يـ	ـيـ	ـي

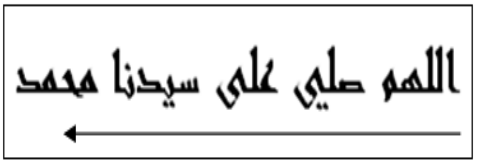
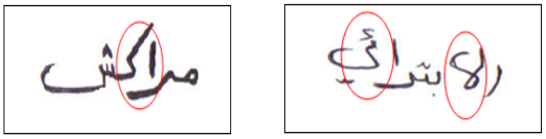


Fig 1. Direction of the Arabic written text

3- The Arabic words may consist of two or more subwords, the words divided into sub words if one of the following (أ, د, ذ, ر, ز, و) characters exist in the middle of the word. Fig 2 shows the Arabic handwriting word consists of five subwords.

Fig 2. Arabic handwriting word 'Alordon' (الاردن) consists of Five sub words

4- Arabic words may contain ligatures and overlapping. The overlapping occurs whenever two or more characters overlap each other, see Fig 3.a. The ligatures occur wherever two or more characters touch each other, see Fig 3.b.



(a) Overlapping (b) ligatures

Fig 3. Arabic handwriting words: (a) Overlapping and (b) ligatures.

5- Arabic words may contain diacritics called Tashkeel, the diacritics are considered as short vowels. These Tashkeels are namely; Fatha, Dhamma, Kasra, Sukun, Madda, Shadda and Tanween. Fig 4 shows the position of these diacritics associated with characters.

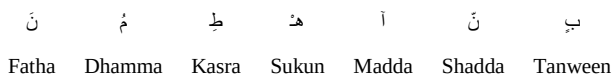
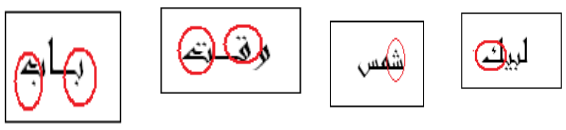


Fig 4. The position of diacritics associated with some characters.

6- Fifteen out of 28 basic Arabic characters have from one to three dots. These dots differentiate one character from another of the same shape, those characters are (ب ,ت ,ث ,ج ,خ ,ذ ,ز ,ش ,ض ,ظ ,غ ,ف ,ق ,ن ,ي), some characters have zigzag shape called hamza such as (أ ,ك). Fig 5 shows some of the Arabic word examples with dots and hamza.



(a) (b) (c) (d)

Fig 5. Examples of Arabic word (a) with one dot (b) two dots (c) three dots (d) with hamza.

7- An Arabic word usually consists of two or more characters which are connected through an imaginary line called the baseline. Fig 6 shows an example of such baseline.

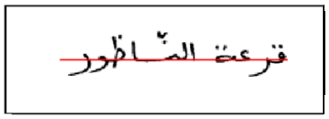


Fig 6. The Arabic baseline

III. OFFLINE ARABIC CHARACTER RECOGNITION

The typical AOCR system consists of five components: image acquisition, preprocessing, segmentation, feature extraction and classification (recognition) [18] [36] [56]. Each of those contributes to the final recognition rate to improve of the OCR system. Fig 7 shows the offline Arabic recognition general frame work. The operation of AOCR will be explained and clarified in the next subsections.

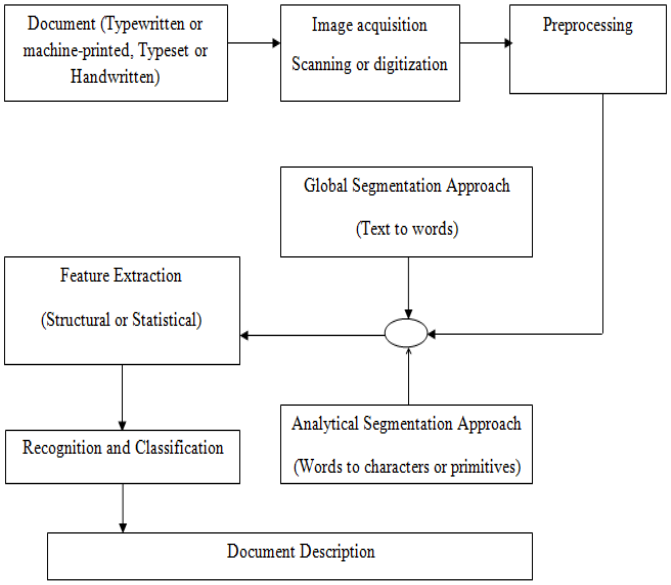


Fig 7. Offline Arabic recognition frame work.

3.1 IMAGE ACQUISITION

Transforming the written text in papers or transcript to digital format is a necessary step in the offline AOCR. In the image acquisition the paper is scanned or captured. The scanner speed, document types and scanning quality need to be considered in the document scanning. Fig 8 shows the image acquisition operation.

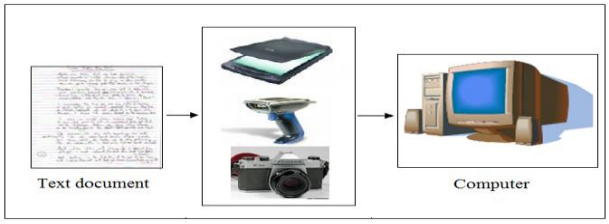


Fig 8. Image question in Arabic OCR system.

3.2 PREPROCESSING

The aims of image preprocessing are to reduce the noise coefficients and to increase the readability of the input by the processing system. The preprocessing stage is also necessary to increase the uniformity in texts which is quite essential for recognition system.

The Preprocessing stage is the most important stage of AOCR. It directly affects the reliability and efficiency in the segmentation, feature extraction and classification process [26] [22]. In order to improve the AOCR system performances, generally, preprocessing stage should contain binarization, filtering and smoothing, slant correction, skew detection, thinning (Skeletonization) and baseline detection, (Fig 9). The operations of AOCR preprocessing are explained and discussed in the next subsections.

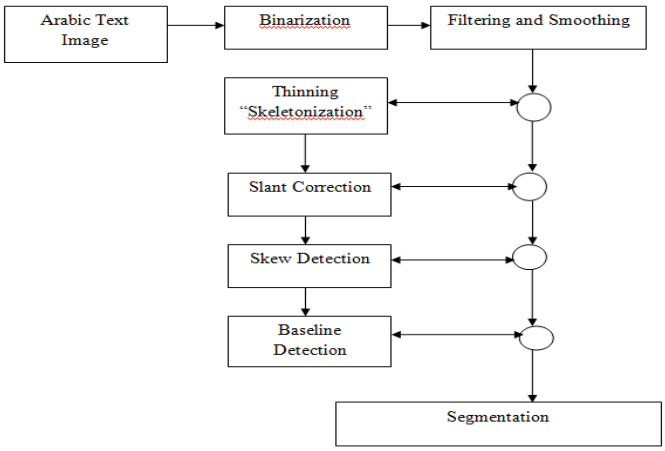


Fig 9. Preprocessing in Arabic optical recognition.

3.2.1 BINARIZATION

The AOCR systems usually accept inputs in bi-level format or to be more specific binary format. Generally, the input text images in grayscale. Hence, we need a preprocessing stage called Binarization. It converts from gray scale image to bi-level image taking into consideration a threshold pixel value for comparison. The threshold pixel value can be computed based on the histogram of the gray values of the images. Fig 10 shows the Arabic handwritten word image "Houda" (هدى) after converted from grayscale to binary format using global threshold.

(a) (b)

Fig 10. Binarization (a) Arabic handwritten word image "Huda" (هدى) In Grayscale format (b) In Binary format

3.2.2 FILTERING AND SMOOTHING:

Noise may appear in the images after scanning or Binarization. It is necessary to remove the noise and smoothing the input text image to prepare the data for the further processing. Generally the AOCR systems are very sensitive to noise. It influences negatively the system performances. The images processing Median or Gaussian filters usually use to remove the noise [20] [13. Fig 11 shows two filtered text images, in the first, the noise was removed by using median filter with window size 3-3, while in the second, the noise was removed by using median filter with window size 5-5.

(a) (b)

Fig 11. Results of removing noise from the original text image using median filter with window size of (a) 3·3 (b) 5·5.

The nature of the Arabic writing or the Binarization process may produce small holes or unwanted edges. These small holes and unwanted information can affect directly in the systems performance. The small holes should be closed and the unwanted information should be deleted by using the opening and closing morphology operation respectively. In Fig 12 the Arabic handwritten word image "Nahal" (نحال) smooth by using closing morphology operation. .

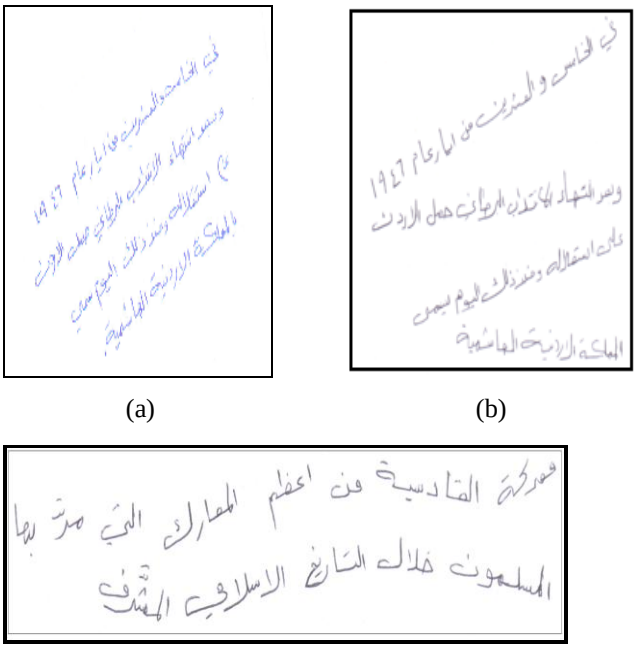
(a) (b)

Fig 12. Smoothing (a) Arabic handwritten word image "Nahal" (نحال): Contains many holes (b) After closing morphology operation

3.2.3 SKEW DETECTION AND CORRECTION

Skew detection and correction is the first step in the document analysis and understanding processing [34]. Correction the skewed image is important in AOCR, because it has a direct effect on the reliability and efficiency of the baseline detection, segmentation and feature extraction [4]. The skew is generally introduced into the image while scanning and leaving it as it is without correction, will give wrong results during document analysis and recognition [4] [34].

The skew detection and correction can be classified into three different categories: First: skewed angle at the paragraph or the document level, this usually occurs because the process of scanned. Second: at the line level. Third: at word level. They often occur because of the nature of the Arabic text writing [49]. Fig 13 shows examples of three different Arabic texts require performing skew detection and correction at the paragraph, line, and word levels.



(a) (b)

(c)

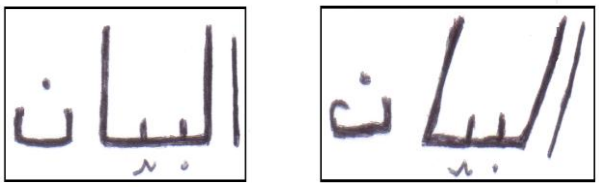
Fig 13. Arabic handwritten texts require performing skew detection and correction at (a) paragraph (b) line, (c) word level.

In AOCR, the Hough transform, Cross Correlation, Projection Profile, Fourier transform and K nearest neighbor (K-NN) clustering can be used for correction the skewed images, such as ([20] [58] [61] [57] [29], More details about the skew detection and correction can be found in [4].

3.2.4 SLANT ANGLE ESTIMATION AND CORRECTION

The slant is a common problem in the Arabic handwriting text images; it occurs because the differences between the text handwriting styles. The slant problem occurs if the vertical components are standing in slant form on the cursive text baseline. It supposes to stand on the perpendicular form and lies above the Arabic text baseline. These vertical components are called Ascenders [3]. More explanation can be found in the Fig 14. Leaving the slanted text without correction, leads to the wrong result in the later stages of the system such as baseline detection and feature extraction.

During the process, the slop of Ascenders must be detected before correction the slant one. The slop of ascenders can be calculated by computing the center of gravity for each stork, and then it can be rotated based on the detected skew angle [47]. The gradient orientation histogram can also use to correct the slant of the handwriting text images [34].



(a) (b)

Fig 14. Slant Estimation (a) the Arabic handwritten word "Al-bayan" (البيان) does not require performing slant correction (b) require performing slant correction.

3.2.5 THINNING OF ARABIC TEXT

Thinning is very important in AOCR system. It simplifies the Arabic texts shapes for segmentation process, feature extraction, and classification. This is resulted in reducing the amount of data that need to be handled [32] [36] [1]. Fig 15 illustrates the impact of thinning in simplifying the Arabic handwritten word shape, in which it shows the word Methlin (مثلين) before and after thinning. The example shows that skeleton perceives the shape of the Arabic word. The Huang et al [35] thinning method is applied on this example and showed that the number of pixel in the word image Methlin (مثلين) was reduced from 4174 pixels to 578 pixels.

(a) (b)

Fig 15. Arabic handwriting word 'methlin' (مثلين): (a) before thinning; (b) after thinning

Many AOCR systems have been developed based on the text skeleton [36] [1]. The skeleton was extensively used in supporting each of feature extraction and classification stages [32]. Beside that its used as the basis for many methods designed for Arabic text segmentation [8], more details about the segmentation methods based on text skeleton can be found in [8]. Other technique based on skeletons includes the estimation the Arabic handwriting word baseline [37].

In the literature, only few thinning algorithms designed especially for the Arabic text. Mahmoud et al [54] proposed a non-iterative thinning algorithm, which was called clustering-based skeletonization algorithm (CBSA). The cluster centres and the adjacency relations between the clusters are used to construct the skeleton of the character. CBSA applies the fuzzy ISODATA clustering algorithm, which consists of iterations until the image is totally clustered. Instead of the fuzzy ISODATA clustering algorithm, Altuwaijri and Bayoumi [39] proposed the Adaptive Resonance Theory (ART2) for clustering of Arabic characters, to speed up CBSA. Another two parallel text thinning algorithms have been proposed by Tellache et al. [41] for AOCR system. The first algorithm works through four sub-iterations, each of which follows a certain procedures. While the second, extracts the skeleton based on the matching between the input text images. Many other thinning algorithms have been proposed [38] [1] [26] [32].

Several thinning algorithms, designed for different purposes, have been used to extract the skeleton of Arabic text, Mostafa [42] used the non-iterative thinning algorithm, which was created by Kegl and Krzyzak [19], to segment the Arabic cursive printed words into characters or into small primitives. In addition, Benouareth et al [9] used the sequential thinning algorithm created by Pavlidis [59] for Arabic handwritten word recognition using Hidden Markov Models with explicit state duration. Many other thinning algorithms have been used to extract the skeleton of Arabic text, such as [46] [21] [12] [60].

3.2.6 BASELINE DETECTION

Detecting Arabic baseline is very important in AOCR because it can be used to segment the Arabic text to characters and make the text ready for the feature extraction stage [5] [8]. Also baseline has been used by most of the AOTR systems [25].

Detecting baseline is one of the main operation in Arabic preprocessing OCR system stage [22] [23], and it is one of the Arabic written characteristic because Arabic language is written cursively [48], the baseline can be used in either skew normalization [37], or for segmenting the text into words or characters [5] [45], also it can be sued to extract dependent features [48]. Using baseline, the characters and shape are classified into three groups Ascenders, descenders and special marks called diacritics such as dots, shadda (Zigzag) and maddah, and these groups may be constructed from stroke or small element or complete character. The Arabic language character shapes based on baseline shown in Fig 16. Ascenders lie above the baseline, but descenders lie under the baseline,

and special marks lie in either above or under the baseline depending on the character [26] [22] [23].

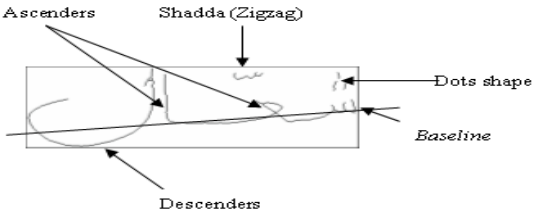


Fig 16. Arabic character shapes based on baseline

AL-Shatnawi & Omar [3] classified the Arabic baseline detection methods into four different groups based on the techniques used. The methods include, baseline detection methods based on horizontal projection, based on word skeleton method, based on contour tracing, and based on principle component analysis. For more details about the methods of Arabic baseline detection and the challenges in detecting Arabic handwritten baseline refer to [3].

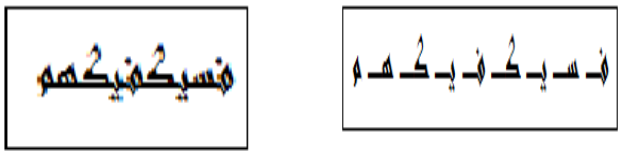
3.3 SEGMENTATION

Segmentation problem is the most difficult and important issue in the AOCR. It directly affects the feature extraction and classification process [49] [8]. Although many segmentation methods have been created for segmenting the Arabic text, the problem is still remain as unsolved issue. This may due to the complex characteristics of the Arabic written text, which have been described earlier in this research [8].

The Arabic text segmentation methods can be classified broadly into two approaches: First is called holistic approach or segmentation-free approach. This technique aims to segment the Arabic text to words or sub words [36] [8]. It splits the paragraph into separate lines and then split these lines into words or sub words. The horizontal projection method is usually used to segment the paragraph into lines [6] while the vertical projection method is used to segment the lines into words or sub words [40]. For more details about the segmentation holistic methods refer to [18]. Second approach is called Analytical approach, in this method, the Arabic word or sub words segment into small classifiable element or into tokens (sliding windows), these elements or tokens could be characters, mixture of characters or strokes [8]. For more details about the analytical segmentation approaches refer to [8]. Two examples about the holistic and analytical approach are shown in the Fig 17 and the Fig 18 respectively. Fig 17 shows the Arabic sentence "Alhoma Sali Ala Sidina Mohammed" (اللهم صلي على سيدنا محمد) before and after holistic segmentation. Fig 18 shows Arabic printed word "Fasikfikohm" (فسيكفيكهم), before and after analytical segmentation.

(a) (b)

Fig 17. Holistic segmentation (a) Arabic sentence "Alhoma Sali Ala Sidina Mohammed" (اللهم صلي على سيدنا محمد) before segmentation (b) after holistic segmentation.



(a) (b)

Fig 18. Analytical segmentation (a) Arabic printed word "Fasikfikohm" (فسيكفيكهم) before segmentation (b) after Analytical segmentation.

Segmentation of Arabic cursive text is still one of the Arabic OCR challenges. These challenges include, ligatures and overlapping, the short distance between two connected characters, and Arabic writing uses many fonts and writing styles.

3.4 FEATURE EXTRACTION:

Feature extraction process is also an important stage in the AOCR system. It has a big influence on the classification stage [36]. The feature extraction process is used to analyze the segmented features of the Arabic text for the classification purposes, and in some cases the combination between several segmented features could enhance the overall recognition rate [17] [43].

The Arabic text feature extraction methods can be classified broadly into three main groups, these groups are structural features, statistical features and Global Transformations: the structural features method, in this technique, features are usually extracted based on the text topologies; the structural features of Arabic text may include loops, the intersection points, dots, zigzags, height, width, number of crossing points, and such [27] [7] [44] [17]. The second is called statistical features methods, these techniques are quick and effective, but may be affected by noise. The statistical features used for Arabic text recognition include: zoning, characteristic loci, crossings and moments [18] [28] [30]. The third is called Global Transformation methods; the Global Transformation aim to shorten the text representation in order to get better results. The global transformations methods used for Arabic text recognition include: horizontal and vertical projections, coding, Hough transform, Gabor transform [24] [50].

3.5 CLASSIFICATION (RECOGNITION):

Classification is the last step in the AOCR system. It assigns an unknown feature into a predefined class. AOCR systems can recognize the text by either the Holistic (Global) or Analytic strategies. The Holistic (Global) strategy recognizes the whole words or sub words, as well as it does not require segmentation, and it works on limited number of vocabularies [11] [52] [49]. On the other hand the Analytic Strategy recognizes the segmented features, as well it requires segmentation, and can be applied on unlimited vocabularies [53] [48]. A number of classification methods are used for Arabic text recognition for examples: Template Matching, Statistical Techniques, Syntactic Techniques, Neural Networks and Hidden Markov Model [6] [11] [49].

IV. DISCUSION

In this paper, the Arabic written text challenges and characteristics of Arabic OCR system are reviewed and studied in details. In addition to what has been presented earlier in this paper, various issues related to the problem of AOCR system and further comments are given in this section. Those issues are discussed in the following points:

- The nature of the input datasets, greatly influence the AOCR system performances. The writing style can be divided into three categories. The first is typewritten (Machine-printed) style, which also called computer-generated. It uses the similar writing style for the whole Arabic characters. The second category is typeset style. The ligatures or the overlapping may appear in this style. The typeset style usually uses to print books, journals, magazines, announcements and newspapers. The third style is handwritten. The size of the vocabulary and the writer dependency effect in this style, which leads to wrong recognition [16].
- The diacritics, such as dots and zigzag, have significant effects on the AOCR system performance in both accuracy and consuming time. Diacritics should be eliminated before any process of AOCR operations. Diacritics elimination may increase the performance time, as well as it may preserve the text topological.
- The development of AOCR systems had not received enough care by researchers, compared with Latin, Chinese and Japanese OCR systems [38]. Where Latin recognition started since 1940 [55], the first attempt to recognize the Arabic language was in 1975 [15].
- The AOCR system consists of five stages: Image acquisition, Pre-processing, Segmentation, Feature Extraction and Classification "Recognition". These stages work together to improve AOCR systems recognition ratio, moreover to reduce the recognition time. Each stage has an impact on the effectiveness and efficiency of the system.
- Preprocessing is the first stage of the AOCR system, it is the most important because it directly affects the reliability and efficiency in the segmentation, feature extraction and classification process. A bad result of this stage produces the wrong input for the other stages of the system. It leads to damage the system, even if the methods were used in these stages are effective and efficient.
- Thinning of Arabic text is one of the main operations in Arabic preprocessing OCR system. It simplifies the text shape and reduces amount of data that need to be handled. The effective thinning algorithm must preserve each of the dots and text connectivity: it also does not produce spurious tails, and it must be robust to noise, as well as it avoids the necking problem.
- Segmentation problem is the most difficult issue in the AOCR, it directly affect each of the feature extraction and classification process [49] [9]. Although many segmentation methods have been created for segmenting the Arabic text, the problem is still an unsettled issue, because of the complex characteristics of the Arabic written text, which is described earlier in this research.
- The feature extraction process is used to analyze the segmented features of the Arabic text for the classification purposes, and in some cases the combination between several segmented features could enhance the overall recognition rate. The classification is the last step in the AOCR system; it assigns an unknown feature into a predefined class.

V. CONCLUSION AND FUTURE DIRECTION

In this paper, the Arabic language characteristics were clarified, and the operations of Arabic offline optical character recognition system stages were discussed and clarified. As well as the operations of AOCR preprocessing stage were also provided and discussed in details. This paper also provides and discuses the challenges that must be considered in designing or choosing a cretin method for the AOCR system. Arabic character recognition is more difficult than the other languages such as Latin or Chinese because the text is written cursively in addition to the complexity of the text characteristics. The nature of the text written need to be considered and studied as a challenge before designing the typical AOCR system, this area is still open for further research such as writer identification and fonts separations. The typical AOCR system consists of five components: image acquisition, preprocessing, segmentation, feature extraction and classification (recognition). Each of those contributes to the final recognition rate to improve of the AOCR. The exiting Arabic OCR systems are still far than the human brain power, in the aspects of accuracy and speed. Each stage has an impact on the effectiveness and efficiency of the system. This study concluded that choosing or designing the effective algorithm for each of the Arabic optical character recognition operation is a crucial issue, when the designed or selected algorithm is ineffective, the system will be negatively affected. Furthermore, the methods proposed in the literature works well

with the printed text, but it works badly with the handwritten text. It is clear that no perfect AOCR system is available yet. Hence, this area of research is still open for further enhancement.

Acknowledgment

The Corresponding Author (Dr. Atallah, M, AL-Shatnawi) would like to thank Dr Houda Bououden for her supports, helps and considerations in carrying out this research.

REFERENCES

[1] A. Ali and Jumari. Skeletonization algorithm for Arabic handwriting, Arab gulf journal of scientific research ISSN 1015-4442, 2004 .vol. 22, no1, pp. 28-33.

[2] A. Alimi and O. Ghorbo. The analysis of error in an on-line recognition systems of Arabic handwritten characters. Proceedings of the Third International Conference on Document Analysis and Recognition. 1995. 2: 890 – 893.

[3] A. AL-Shatnawi, and K. Omar, Methods of Arabic Baseline Detection - The State of Art, International Journal of Computer Science and Network Security. 2008. Vol.8 (10):137-142 .October.

[4] A. AL-Shatnawi, and K. Omar, Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity. Journal of Computer Science, ISSN 1549-3636, 2009. vol.5 (5): 363-368.

[5] A. Amin, "Offline Arabic Character Recognition: The State of the Art," Pattern Recognition, 1998. vol. 31, pp. 517-530.

[6] A. Amin, and J.F. Mari, Machine recognition and correction of printed Arabic text. IEEE Transactions on Systems, Man and Cybernetics (SMC), 1989. 19(5): 1300- 1306.

[7] A. Amin and H. Alsadoun, Hand printed Arabic character recognition system. Proceedings of the 12th International Conference A on Pattern Recognition, IAPR ; 1994. pp 536–539.

[8] A.M. Zeki, The segmentation problem on Arabic character recognition – the state of the art. 1st International Conference on Information and Communication Technology (ICICT). 2005. pp. 11-26. Karachi, Pakistan.

[9] A. Benouareth, A .Ennaji, and M. Sellami. Arabic handwritten word recognition using HMMs with explicit state duration. EURASIP Journal on Advances in Signal Processing. 2008 (1-13).

[10] A. Broumandnia, J. Shanbehzadeh, and M. Nourani3, Handwritten Farsi/Arabic Word Recognition. IEEE. 2007. pp. 767-771.

[11] A. Dehghani, F .Shabani and P. Nava, Off-Line Recognition of Isolated Persian Handwritten Characters Using Multiple HiddenMarkov Models, Proc. Int'l Conf. Information Technology: Coding and Computing, 2001. pp. 506-510.

[12] A .Ferreira, and S. Ubeda Ultra fast parallel countor tracking with application to thinning. Pattern Recognition, 1994. 27(7):867–878.

[13] A .Gross, and L. Latecki, Digital geometric methods in document image analysis. Pattern Recognition.1999. 32(3), pp. 407.

[14] A. M. Sarhan and O, I. Al Helalat, .Arabic character recognition using artificial neural networks and statistical analysis. Proceedings of world academy of science, engineering and technology. ISSN 20071307-6884. 21: 32-36. May.

[15] A. Nazif, A system for the recognition of the printed Arabic characters. M.Sc. Thesis. Cairo University. 1975.

[16] A. Zaki, Segmentation of Arabic Characters Using Voronoi Diagrams. PhD. Thesis. Faculty of Information Science and Technology, University Kebangsaan Malaysia, Malaysia. May. 2008.

[17] A., Zidouri, S Chinveeraphan,., and M. Sato, Structural Features by MCR Expression Applied to Printed Arabic Character Recognition. in 8th Int. Conf. on Image Analysis and Processing, (San Remo Italy), 1995.pp.557--562, Sept. 13-15

[18] B. AL -Badr and S. Mahmoud . Survey and bibliography of Arabic optical text recognition. Signal Processing. 1995. 41(1): 49-77.

[19] B. Kegl, and A. Krzyzak, Piecewise linear skeletonization using principal curves. IEEE Transactions on Pattern Analysis and Machine Intelligence 2002. 24, N 1, pp. 59-74, 2002.

[20] C. Gonzales. Rafael and E. Richard ,Woods. Digital Image Processing. 2nd ed. Englewood Cliffs, NJ: rentice-Hall. 2002.

[21] C. J. Hilditch, Comparison of thinning algorithms on a parallel processor. Image Vision Computing, 1983. 1, 115–132.

[22] F. Farooq, V .Govindaraju, and M. Perrone, Pre-processing Methods for Handwritten Arabic Documents. (ICDAR'05) Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition, IEEE. 1. 2005. pp. 267-271.

[23] F. Latfi, F .Nader, and B. Mouldi, Arabic Word Recognition by Using Fuzzy Classifier). Journal of Applied Scinces. ISSN 2006. 1812-56546. (3): 617-650.

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

[25] H. Almuallim, and S. Yamaguchi. A method of recognition of Arabic cursive handwriting. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 1987. 9(5): 715-722.

[26] H. Al-Rashaideh, Preprocessing phase for Arabic Word Handwritten Recognition. Russian Academy of Sciences. 2006. 6(1): 11-19. Russian Federation.

[27] H. Goraine, M. Usher, and S. Al-Emami, "Off-Line Arabic Character Recognition," Computer, 1992. vol. 25, pp. 71-74.

[28] H. Sanossian, An Arabic character recognition system using neural network. Proceedings of 1996 IEEE Signal Processing Society Workshop, Kyoto, Japan, IEEE1996., pp; 340–348.

[29] H. Yan, Skew correction of document images using interline cross correlation, Computer Vision, Graphics, and Image Processing 55, 1993. pp 538-543. Doi: 10.1006/cgip.1993.1041.

[30] I. Bazzi, R .Schwartz, and J. Makhoul, An omnifont open-vocabulary ocr system for English and Arabic. IEEE Trans Pattern Analysis and Machine Intelligence. 1999. vol; 21(6): 495–504

[31] J, H. AlKhateeb J. Ren S . Ipson and J. Jiang, knowledge-based baseline detection and optimal thresholding for words segmentation in efficient pre-processing of handwritten Arabic text. Fifth international conference on information technology: new generations. IEEE computer society. 2008. pp. 1158-1159.

[32] J. Cowell, and F. Hussain, Thinning Arabic characters for feature extraction. IEEE Conference on Information Visualization. London, UK. 2001. pp. 181-185. 25-27 July.

[33] K .Jumari and M, A. Ali, A Survey And Comparative Evaluation Of Selected Off-Line Arabic Handwritten Character Recognition Systems. Jurnal Teknologi, 2002. 36(1-18) Jun.

[34] K. Omar, A. Ramli , R. Mahmod, and M. Sulaiman, Skew Detection and Correction of Jawi Images Using Gradient Direction. Journal Technology, 2002. VOL 37(D) 117–126. Dis.

[35] L. Huang, G. Wan, and C. Liu, An Improved Parallel Thinning Algorithm. Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), 2003. 780-783.

[36] M .Liana, and G. Venu, Offline Arabic Handwriting Recognition: A Survey. IEEE, Transactions on Pattern Analysis and Machine Intelligence. 2006. 28: 712-724.

[37] M .Pechwitz, and V. Maergner, Baseline estimation for arabic handwritten words. In Frontiers in Handwriting Recognition. 2002. 479–484.

[38] M. Melhi. S. Ipson. and W. Booth. A novel triangulation procedure for thinning hand-written text, Pattern Recognition Letters, 2001. (22)1059–1071.

[39] M. Altuwaijri, M. Bayoumi, A thinning algorithm for Arabic characters using ART2 neural network. CirSysSignal (45), no. 2, February, 1998. pp. 260-264.

[40] M. Sarfraz, S.N. Nawaz, and A. Al-KhuraidlyOffline Arabic text recognition system. International Conference on Geometric Modeling and Graphics (GMAG'03). , 2003. pp. 30-36. London, England. 16-18 July.

[41] M., Tellache, M. Sid-Ahmed, and B. Abaza, Thinning algorithms for Arabic OCR. IEEE Pacific Rim Conference on Communications and Signal Processing. 11993.: 248-251. Victoria, BC, USA. 19-24 May.

[42] M.G. Mostafa, An adaptive algorithm for the automatic segmentation of printed Arabic text. 17th National Computer Conference. 2004. pp. 437-444. Madinah, Saudi Arabia. 5-8 April.

[43] M.S Khorsheed, and W.F. Clocksin,. Segmentation-free word recognition for Arabic handwriting. The International Conference on Pattern Recognition ICPR'2000, Barcelona, Spain, September.

[44] M.S Khorsheed. and W.F. Clocksin, "Structural Features of Cursive Arabic Script," Proc. British Machine Vision Conf., 1999. pp. 422-431.

[45] N .Arica, and F. Yarman-Vural. Optical character recognition for cursive handwriting, IEEE PAMI. 2002 .24 (6):801 – 813.

[46] N. J Naccache. and R. Shinghal. SPTA: A Proposed Algorithm for Digital Pictures. IEEE Trans. on Systems, Man and Cybernetics, 1984. vol. SMC-14(3) 409-418.

[47] R. Bozinovice, and S. N. Srihari. Off-line cursive word recognition IEEE Trans. On PAMI. 1989. 11: 68 – 83.

[48] R. El-Hajj, L .likforman-Sulem, and C. Mokbe, Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling. (ICDAR'05) Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition, IEEE. 20 (5). 2005. pp. 1520-5263.

[49] R. Safabakhsh, and P. Adibi, Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM. The Arabian Journal for Science and Engineering. 2005. 30: 95-118. April.

[50] S .Saadallah, and S. Yagu, Design of an arabic character reading machine. Workshop on Computer Processing and Transmission of the Arabic Language, Kuwait. 1985.

[51] S. Al-Emami and M. Usher. On-line Recognition of Handwritten Arabic Characters. IEEE Trans. Patt Anal. Machine Intell. 1990. 12(7): 704 – 710.

[52] S. Alma'adeed, C. Higgens. and D. Elliman, Off-Line Recognition of Handwritten Arabic Words Using Multiple Hidden Markov Models, Knowledge-Based Systems, 2004. vol. 17, pp. 75-79.

[53] S. Mozaffari, K. Faez, and M. Ziaratban, Structural Decomposition and Statistical Description of Farsi/Arabic Handwritten Numeric Characters, Proc. Int'l Conf. Document Analysis and Recognition, 2005. pp. 237-241.

[54] S., Mahmoud, I. Abuhaiba, and R. Green, Skeletonization of Arabic characters using clustering based skeletonization algorithm (CBSA). Pattern Recognition. 1991. 24(5): 453-464.

[55] S.A. Alshebeili, A.A. Nabawi, and S.A. Mahmoud, Arabic character recognition using 1-D slices of the character spectrum. Signal Processing. 1997. 56(1): 59-75.

[56] S.N Nawaz, M Sarfraz, A. Zidouri, and W.G. Al-Khatib, An approach to offline Arabic character recognition using neural networks. 10th IEEE International Conference on Electronics, Circuits and Systems (ICECS'03). 2003. 3:1328-1331. 14-17 December.

[57] S.N Srihari and V. Govindaraju Analysis of textual images using the Hough Transform, Machine Vision and Applications, 1989.vol 2, pp.141-153. Doi: 10.1007/BF01212455.

[58] T. Akiyama and N. Hagita. Automated entry system for printed documents, Pattern Recognition. 1990. Vol. 23, No. 11, pp 1141-1158.

[59] T. Pavlidis, Algorithms for Graphic and Image Processing, Computer science press, Rockville, Md, USA1982..

[60] T. Y Zhang. and C. Y. Suen. A Fast Parallel Algorithm for Thinning Digital Patterns. Comm. ACM, 1984. vol. 27(3) 236-239.

[61] V .Argner, and H. El Abed, Databases and Competitions: Strategies to Improve Arabic Recognition Systems. 2008. pp. 82-103.

[62] W. Postl, Detection of linear oblique structures and skew scan in digitized documents. Proceedings 8th International Conference on Pattern Recognition, 1986. pp. 687-689.

AUTHORS PROFILE

Atallah. M AL-Shatnawi has received his BSc in Computer Science from Yarmouk University (Jordan) in 2005, MSc in Computer Science from the University Science Malaysia (USM) and PhD in System Sciences and Management from the National University of Malaysia (UKM) in 2007 and 2010 respectively. Currently, he is an Assistant Professor at the Department of Computer Science, Faculty of Science and Information Technology, Irbid National University (Jordan). His research interests include Pattern Recognition, Image Processing and Embedded Systems. He has published numerous papers related to these areas.

Farah Al-Zawaideh is the Chairman of Computer Information System in Irbid National University from 2009 until now. He received a Ph.D. in Knowledge based systems from the Faculty of Computer Information Systems, University of Banking and Financial Sciences. His research interest in Classification using genetic algorithms, E-learning and software engineering. He has a wealth of expertise gained from his work experiences in Jordan, ranging from web development to network administration.

Safwan AL-Salaimeh is currently the vice dean of faculty of Science and Information Technology at Irbid National University. He received his MSc in Electronic Engineering from and PhD in Computer Engineering from Kharkiv State University Ukraine in 1997 and 2001 respectively. His research interests include operation research, information logistics system, and system analysis.

Khairuddin Omar is currently a professor at the Faculty information technology and sciences of Universiti Kebangsaan Malaysia (UKM). He received his MSc in Computer Science from UKM and PhD in Computer Science from Universiti Putra Malaysia in 1989 and 2000 respectively. His research interests include Pattern Recognition in decision making with uncertainty, and Image Processing. He has published numerous papers related to these areas. He currently leads the Pattern Recognition research group at UKM.