Methods of Arabic Language Baseline Detection – The State of Art

Atallah AL-Shatnawi and Khairuddin Omar

University Kebangsaan Malaysia, Faculty of Information Science And Technology, Malaysia

Summary
Preprocessing is the most important stage in the Arabic OCR system; it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages. It is worth mentioning that Arabic language is cursively written, and its characters have between 2 to 4 shapes. An Arabic word likely consists of two or more characters which are connected through an imaginary line called baseline. Detecting baseline is one of the main majorities in preprocessing Arabic OCR system. The baseline can be used for both skew normalization and character segmentation. This paper aims to provide a comprehensive review of the methods proposed by researchers to detect Arabic baseline. The Arabic baseline detection methods are categorized into four methods: (a) based on horizontal projection methods, (b) based on word skeleton method, (c) based on contour tracing method, and (d) based on principle component analysis method. Each of these methods has its own advantages and drawbacks.

Key words:
Preprocessing, OCR, Handwritten, Offline, Arabic Baseline.

Introduction

The goal of the character recognition 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 [11] [34]. In pattern recognition field, languages recognition is considered as one of the most complicated problem in Artificial Intelligent field [1] [32]. Recognition can be done offline or online. In offline recognition, papers, manuscripts or documents are scanned or captured, and finally are manipulated by OCR system. In online recognition application takes place during the writing process [2] [18] [20] [39].

Arabic language is universal and it is a formal language for 25 countries, of population over than 250 million [3] [18] [20]. Additionally, many Arabic characters are used in different languages such as Ardu, Farsi, Chawi, Kardi [39]. Literature is rich with evidence that Arabic language recognition is more difficult than other languages such as Latin or Chinese because the text is written cursively in addition to the complexity of the text characteristics [3] [7] [13], More details about Arabic written characteristics can be founded in each of [2] [18] [20] [39]. The development of Arabic OCR systems had not received enough care by researchers, compared with Latin or china's OCR systems [39]. Where Latin recognition started since 1940 [6], the first attempt to recognize the Arabic language was in 1975 [23]. The Arabic OCR system goes through five stages: Image acquisition, Preprocessing, Segmentation, Feature Extraction and Classification (Recognition) [2] [18] [20] [22]. These stages work together to improve OCR systems recognition ratio moreover to reduce the recognition time [32] [34].

OCR preprocessing stage is the most important because it directly affects the reliability and efficiency in the segmentation and feature extraction process [5] [17]. In order to improve the OCR system performances, preprocessing stage should contain smoothing, noise removal, image decomposition, skew detection and correction, edge detection, and baseline detection, the baseline detection is that research's focus. Detecting Arabic baseline is very important in Arabic OCR because it can be used to segment the Arabic text to characters and make the text ready for the feature extraction stage [8] [39]. Also baseline has been used by most of the OCR systems [4].

Detecting baseline is one of the main majority in preprocessing Arabic OCR system stage [17] [19], and it is one of the Arabic language characteristic because Arabic language is written cursively [15], the baseline can be used in either skew normalization [28], or for segmenting the text into words or characters [8] [12], also it can be sued to extract dependent features [15]. 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 (1). 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 [5] [17] [19].

Manuscript received October 5, 2008
Manuscript revised October 20, 2008

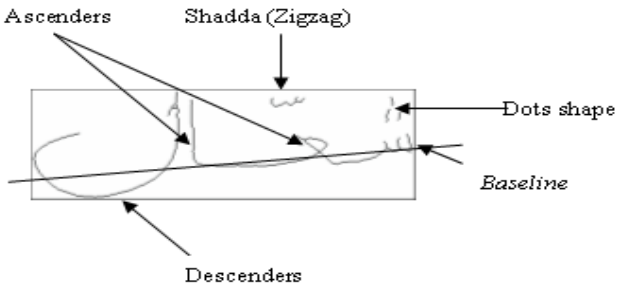

Fig. 1 Arabic language character shapes based on baseline.

The diacritics such as dots, zigzag, and diacritics have significant effects on the OCR system performance [39]. Diacritics should be eliminated in the process of baseline detection when using horizontal projection or principle component analysis techniques because of the bad influences of the detection method performances [5] [14], On the other hand diacritics have no influences on the performance of baseline detection, using word skeleton [29], also sometimes they have influences on the baseline detection method based on the word contour performance [17]. More details about diacritics isolation and their influences on the OCR systems can be founded in each of [7] [16] [24] [26]. Arabic baseline detecting methods have received more concern researchers in the last three decades where the first attempt to detect Arabic baseline used the horizontal projection was in 1981 by Parhami and Taraghi [27], then it improved by Timsari & Fahimi [37], Then the detecting methods sophisticated by the researchers by using different techniques such as contour representation [17], word skeleton [29] and principle component analysis [14].

In this paper, baseline detecting methods are categorized based on the techniques used. In the next sub-sections these techniques will be discussed in detail, including their advantages and drawbacks.

A- Baseline detection Methods Based on Horizontal Projection:

The horizontal projection method is commonly used by the OCR researchers to detect Arabic baseline [29], the first time the horizontal projection method was used to detect Arabic baseline was in 1981 [27]. The horizontal projection method to detect Arabic baseline works by reducing the 2D of data to 1D based on the pixels of the text image, and the longest pink which will implement in the straight line will be the Arabic text baseline. Fig 2: Horizontal projection method for detecting Arabic baseline.

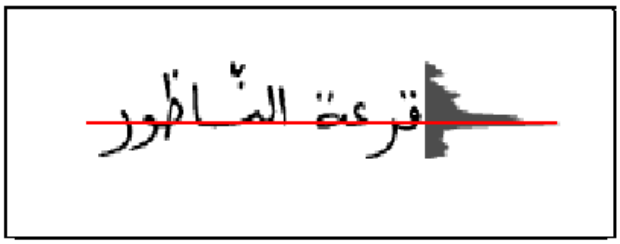

Fig. 2 horizontal projection method for detecting Arabic baseline.

The Horizontal projection profile is defined as:

$$P(i) = \sum Image\ (i,j)$$

Where P(i) is the horizontal projection of the image for row i, and the Image (i, j) is the pixel value at (i, j).

Thickness of baseline is found by calculating the thickness of the greatest spike, using the most frequent column-height [37], or taking into account the location of iteration as a reference since they are often near close to the baseline [25]. Nawaz et al [22] and Sarfraz et al [33], used the horizontal projection method to divide the image into four different zones, baseline, middle, upper and lower zone, shown in fig (3) baseline, middle, upper and lower zone, the baseline zone is of denser black pixels; the middle zone is just above the baseline and double thickness of it.

Fig .3 baseline, middle, upper and lower zone

Ramy El-Hajj et al [15] used the horizontal projection method to detect two baselines in each input image, upper and lower baseline, fig (4) Upper and lower baselines, the lower baseline is detected based on the classical horizontal projection according to the longest peak, and the upper baseline is detected according to the longest peak above the lower baseline.

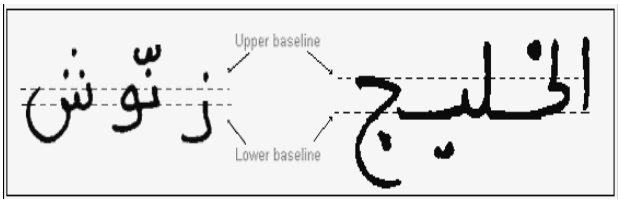

Fig. 4 Upper and lower baselines

AL-Rashidah [5] proposed a more sophisticated way that depends on the iteration with angle to detect Arabic baseline with the horizontal projection, this way is summarized in the following three steps:

1- Apply the horizontal projection profiles for the input image and calculate the highest peak.
2- Rotate the image with constant or static angle, and apply the horizontal projection profiles and find the highest peak.
3- Compare the two peaks values if the value in the second step is higher than the first, apply the second step again, else stop and determine the baseline.

AlKhateeb et al [3] proposed to divide the handwriting word image by middle line, and he supposed that Arabic baseline always lies below the middle line of the image, this leads to horizontal projection of the lower half of an image gives better results than the whole image. The new results using this method are shown in fig (5).

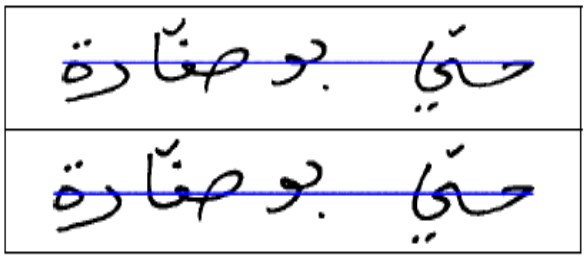

Fig. 5 in the top image the failure baseline detected by classical horizontal projection and in the bottom the improved result using Alkhateeb proposed algorithm.

The horizontal projection of Arabic baseline detection methods is easy to use, most common and it works well with the printed text [29] [39]. Yet it is difficult to detect Arabic baseline from the handwriting text especially with handwriting word, because it detects the baseline in the straight lines [29]. Moreover the diacritics are still one of it is challenges, because they affect drastically methods performances. The horizontal projection profile of Arabic baseline detection methods had been applied for segmentation [4] [7] [9] [10] [21] [31] [35], skew detection [5] [38], and to extract dependent features [15]. Segmentation was based on thickness of the connection point between the stroke and the baseline or between the strokes, because there is more intensity of the thickness between other parts of the word than of the thickness between the stroke and the baseline. For skew detection it was based on the text writing nature for handwriting text. Only one study had applied extracting dependent features [15], where the image segmented into slides according to the detected baseline, and then from the segmented slides the dependent features were extracted.

B- Baseline Detection Method Based on The Word Skeleton:

Only one work had been proposed to detect Arabic handwriting baseline according to the word skeleton [29]. This method based on the skeleton lines of the polygonal word skeleton. This work is summarized in the following four steps:

1- The polygonal approximations create the skeleton of the word.
2- Set of the feature extractions calculated, based on the polygonal skeleton, and these features represent the baseline relevant features.
3- The best extracted baseline relevant features will be chosen according to a set of conditions listed, depending on the Arabic writing style.
4- The baseline detected by using linear regression depending on the selected relevant features which will convert to the baseline estimate point in the previous step.

This method works well with the Arabic handwriting words, also it can be applied in printed text, and it is more flexible with the different word handwriting styles, the diacritics do not affect the method performance, on the other hand it needs more time than other baseline detection method because it works with set of the complex calculations. Fig (6) shows the Arabic baseline estimation for Arabic handwriting word based on the word skeleton.

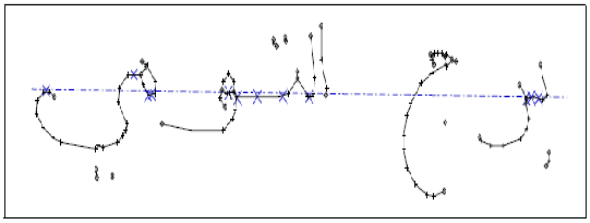

Fig .6 estimation Arabic baseline based on the word skeleton

C- Baseline Detection Method Based on Word Contour Representation:

Only one work had been proposed to detect Arabic handwriting baseline according to the word Contour representation [17]. This method based on local minima points of the word contour. This work is summarized in the following four steps:

1- Locate the local minima points from the word counter. The minima points depend on the contour direction; it is located where the contour changes direction from the down word to the up word.
2- Apply (least squared sum) linear regression to detect the approximate baseline.
3- Reset anew minima points close to the old point.
4- Use the second linear regression to detect the word baseline.

This method works well with the Arabic handwriting word. Also it can be applied in the printed text, and it is more flexible with the different word handwriting styles, it can work with or without diacritics, but they are affected in the method performances if their size large relative to the word. Fig (7) shows the Arabic baseline estimation for two words based on the word contour representation.

Fig .7 estimation Arabic baseline based on the word contour representation

D- Baseline Detection Method Based on The Principal Components Analysis (PCA):

Principal Components Analysis (PCA) can be used to compress and recognize the two domination and the three domination images as well as it used for Latin skew detection [36] and Arabic baseline detection [14]. Only one work had been done to detect Arabic handwriting baseline based on the principle Components Analysis [14], this method based on angle detected by principle Components Analysis. It is used to determine the Arabic word baseline direction according to the foreground or background pixels distribution. This work summarized in the following four steps:

1- Perform the Principal Components Analysis for either foreground or background pixels distribution.
2- Determine baseline estimate direction by using PCA eigenvector.
3- Rotate the images according to the estimate direction and calculate the rotation angles.
4- Apply Horizontal projection to determine the longest pink, and detect the vertical position of the baseline.

This method works with both foreground or background pixels, and it gives good result with background pixels more than foreground [36]. Also it works with both diacritics or without diacritics, but it works better without diacritics [14]. Also it can be applied in the printed text. Fig (8) shows the Arabic baseline estimation for two words based on the principal components analysis.

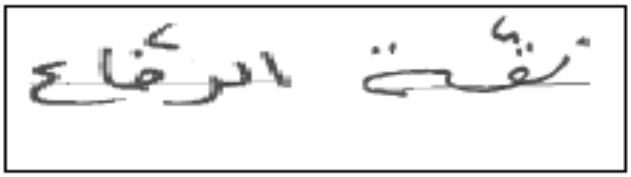

Fig. 8 estimation Arabic baseline based on the principal components analysis

This method is based on Angle like that sophisticated horizontal projection one which proposed by AL-Rashidah [5], but this method find the angle in one step while the second method which relay on the image rotation.

List of Difficulties in Detecting Arabic Handwriting Baseline

Detecting Arabic baseline is still one of the Arabic OCR challenges, these challenges include the following:

1- The first challenge is that Arabic is written cursively, and it has 28 characters and each character written between two to four ships according to its location in the word, see table 1 Shapes of Arabic characters in different positions. Additionally, Arabic words may contain ligatures and overlapping [39]. Fig 9 overlapping and ligatures with Arabic words.

Character Name		Final	Medial	Initial	Isolated
Alif	ألف	ـا	ـا	ا	ا
Ba'	باء	ـب	ـبـ	بـ	ب
Ta'	تاء	ـت	ـتـ	تـ	ت
Tha'	ثاء	ـث	ـثـ	ثـ	ث
Jeem	جيم	ـج	ـجـ	جـ	ج
H'a'	حاء	ـح	ـحـ	حـ	ح
Kha'	خاء	ـخ	ـخـ	خـ	خ
Dal	دال	ـد	ـد	د	د
Th'al	ذال	ـذ	ـذ	ذ	ذ
Rai	راي	ـر	ـر	ر	ر
Zai	زاي	ـز	ـز	ز	ز
Seen	سين	ـس	ـسـ	سـ	س
Sheen	شين	ـش	ـشـ	شـ	ش
S'ad	صاد	ـص	ـصـ	صـ	ص
Dhad	ضاد	ـض	ـضـ	ضـ	ض
T'a'	طاء	ـط	ـطـ	طـ	ط
Dh'a'	ظاء	ـظ	ـظـ	ظـ	ظ
'Ain	عين	ـع	ـعـ	عـ	ع
Ghain	غين	ـغ	ـغـ	غـ	غ
Fa'	فاء	ـف	ـفـ	فـ	ف
Qaf	قاف	ـق	ـقـ	قـ	ق
Kaf	كاف	ـك	ـكـ	كـ	ك
Lam	لام	ـل	ـلـ	لـ	ل
Meem	ميم	ـم	ـمـ	مـ	م
Noon	نون	ـن	ـنـ	نـ	ن
Ha'	هاء	ـه	ـهـ	هـ	ه
Waw	واو	ـو	ـو	و	و
Ya'	ياء	ـي	ـيـ	يـ	ي

Table 1 Shapes of Arabic characters in different positions

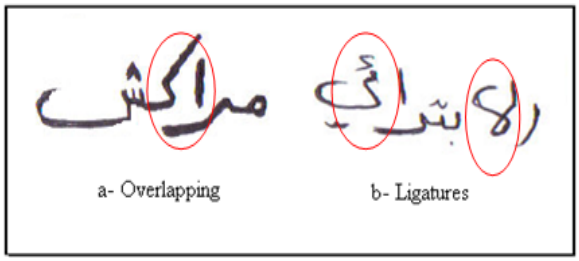

Fig. 9 (a) overlapping and (b) ligatures in Arabic words

2- The diacritics, such as dots and zigzag, have significant effects on the Arabic baseline detection performance in both accuracy and consuming time. Diacritics should be eliminated in the process of baseline detection and this elimination will increase the consuming time.

3- Word slope is also affected on the baseline methods performances. Some researchers manipulate it by finding the baseline direction before starting detecting. This will also increase the consuming time.

4- Finding the Arabic baseline in straight line is still one of the Arabic baseline detection challenges because some Arabic words may be contracted from two or more sub words, and the distribution of these sub words in the same word makes detecting baseline difficult. Figure 10 Arabic handwriting words constructed many sub words.

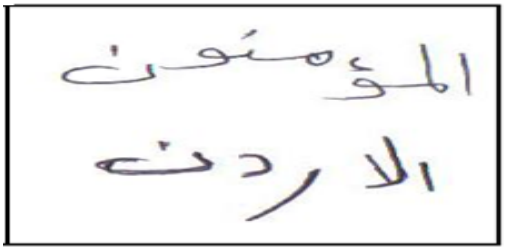

Fig. 10 Arabic handwriting words constructed from many sub words

Future Direction and Conclusion

In this literature, the Arabic baseline detection methods are discussed in detail. The methods proposed in the literatures were classified into four different groups based on the techniques used. And the methods drawbacks and advantages were discussed in detail.

This study concluded that the methods proposed in the literature works well with printed text, but there is a need to develop Arabic handwriting baseline detection methods by using hybrid detection methods that combine between two or more methods or by finding new other techniques work without angle and not effect by the diacritics. From the above review, it seems that only the IFN/ENIT dataset [30], which contains the Tunisian cities name, was used to implement and test the different Arabic word handwriting baseline detection techniques.

This paper addressed the Arabic baseline detection methods; every method is explained according to the general Arabic baseline detection. This paper is the first survey to describe the methods of Arabic baseline detection since 28 years of work. As such, it is clear that no perfect and Arabic baseline detection methods are available yet. Hence, this area of research is still open for further enhancement.

Acknowledgments

Authors would like to thank Ayman Jaradat and Mohammad Nasrudin for their supports, helps and considerations in carrying out this research.

References

[1] Aburas, A, A. & Rehiel, M, A. 2007. Off-line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression. Arab Research Institute in Sciences & Engineering. ISSN 1994-3253. 3(4): 123-135.

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

[3] AlKhateeb, J, H. Ren, J. Ipson, S & Jiang, J. 2008. 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. pp. 1158-1159.

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

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

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

[7] Al-Yousefi, H. & Udpa, S.S. 1992. Recognition of Arabic characters. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 14(8): 853-857.

[8] Amin, A. 1998. Off-line Arabic character recognition: the state of the art, Pattern Recognition. 31(5): 517-530.

[9] Amin, A. 1997. Arabic character recognition. In Bunke H. & Wang P.S.P. (ed.) Handbook of Character Recognition and Document Image Analysis. pp. 397- 420. World Scientific, Singapore.

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

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

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

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

[14] Burrow, P. 2004. Arabic handwriting recognition. M.Sc. Thesis. University of Edinburgh. England.

[15] El-Hajj, R. likforman-Sulem, L & Mokbe, C. 2005. 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). pp. 1520-5263.

[16] Fahmy, M.M.M. & El-Messiry, H. 2001. Automatic recognition of typewritten Arabic characters using Zernike moments as a feature extractor. Journal of Studies in Informatics and Control. 10(3):48-51.

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

[18] Khorsheed, M.S. 2002. Off-line Arabic character recognition - a review. Pattern Analysis & Applications. 5(1): 31-45.

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

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

[21] Najoua, A & Noureddine, E. 1995. A robust approach for Arabic printed character segmentation. In ICDAR. pp. 865–868.

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

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

[24] Nouh, A, Ula, A.N. & Edlin, A.S. 1988. Algorithms for feature extraction: a case study for the Arabic character recognition. 10th National Conference. pp. 653- 666. Jeddah, Saudi Arabia.

[25] Olivier, C., Miled, H., Romeo-Pakker, K. & Lecourtier Y. 1996. Segmentation and coding of Arabic handwritten words. International Conference on Pattern Recognition (ICPR'96). 3: 264-268. Austria.

[26] Omar, K., Mahmoud, R., Sulaiman, M.N. & Ramli, A. 2000. The removal of secondaries of Jawi characters. IEEE Region 10 Annual Conference (TENCON'2000). 2: 149-152. Malaysia.19-22 August.

[27] Parhami, B. & Taraghi, M. 1981. Automatic recognition of printed Farsi texts. Pattern Recognition. 14(1-6): 395-403.

[28] Pechwitz, M & Maegner, V. 2003. HMM Based approach for handwritten Arabic Word Recognition Using the IFN/ENIT– DataBase, ICDAR'03, Edinburgh. pp. 890-894.

[29] Pechwitz, M & Maergner, V. 2002. Baseline estimation for arabic handwritten words. In Frontiers in Handwriting Recognition. 479–484.
[30] Pechwitz, M., Maddouri, S.S., Märgner, V., Ellouze, N. & Amiri, H. 2002. IFN/ENIT - Database of Handwritten Arabic Words. Colloque International Francophone sur l'Écrit et le Document (CIFED'02). pp. 129-136. Hammamet, Tunisia. 21-23 October.
[31] Romeo-Pakker, K., Miled, H. & Lecourtier, Y. 1995. A new approach for Latin/Arabic character segmentation. 3rd International Conference on Document Analysis and Recognition (ICDAR'95). 2: 874-877. Montreal, Canada.
[32] Safabakhsh, R & Adibi, P. 2005. Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM. The Arabian Journal for Science and Engineering. 30: 95-118. April.
[33] Sarfraz, M., Nawaz, S.N. & Al-Khuraidly, A. 2003. Offline Arabic text recognition system. International Conference on Geometric Modeling and Graphics (GMAG'03). pp. 30-36. London, England. 16-18 July.
[34] Sarhan, A, M & Al Helalat, O, I. 2007 .Arabic character recognition using artificial neural networks and statistical analysis. Proceedings of world academy of science, engineering and technology. ISSN 1307-6884. 21: 32-36. May.
[35] Shoukry, A. 1990. A sequential algorithm for the segmentation of typewritten arabic digitized text. The Arabian Journal for Science and Engineering. 16(4b):543–556. Oct.
[36] Steinherz, T. Intrator, N & Rivlin, E. 1999. Skew detection via principal components analysis. In Proceedings of the 5th International Conference on Document Analysis and Recognition, IEEE. pp. 153–156.
[37] Timsari, B. & Fahimi, H. 1996. Morphological approach to character recognition in machine-printed Persian words. SPIE Document Recognition III. San Jose, CA.
[38] Zahour, A., Taconet, B., Mercy, P. & Ramdane, S. 2001. Arabic hand-written text-line extraction. 6th International Conference on Document Analysis and Recognition (ICDAR'01). pp. 281-285. Washington, USA. 10-13 September.
[39] Zeki, A.M. 2005. The segmentation problem on Arabic character recognition – the state of the art. 1st International Conference on Information and Communication Technology (ICICT). pp. 11-26. Karachi, Pakistan.

Atallah. M AL-Shatnawi, received the BA degree in computer science from Yarmouk University in Jordan in 2005, and M.S. degree in computer science from University science Malaysia in 2007, and currently pursuing his PhD in the field of Arabic Character Recognition at the University Kebangsaan Malaysia.

Khairuddin. Bin. Omar, received the BA and the M.S degree in Faculty of Information Science And Technology at the University Kebangsaan Malaysia in 1986 and 1989, and PhD degree in computer science from University Putra Malaysia in 2000. Dr Omar currently is Assoic. Prof in Faculty of Information Science And Technology at the University Kebangsaan Malaysia. Dr Omar research is focused on pattern recognition applications in the areas of biometrics and digital libraries.