Neural-Based Segmentation Technique for Arabic Handwriting Scripts

Husam A. Al Hamad
College of Computer
Qassim University
Saudi Arabia
hushamad@yahoo.com, hhamad@qu.edu.sa

ABSTRACT

In some algorithms, segmentation of the word image considers the first step of the recognition processes; the main aim of this paper is proposed new fusion equations for improving the segmentation of word image. The technique that has used is divided into two phases; at the beginning, applying the Arabic Heuristic Segmenter (AHS), AHS uses the shape features of the word image, it employs three features, remove the punctuation marks (dots), ligature detection, and finally average character width, the goal of this technique is placed the Prospective Segmentation Points (PSP) in the whole parts of the word image. As a result, the second phase apply the neural-based segmentation technique, the goal of neural technique is check and examine all PSPs in the word image in order to report which one is valid or invalid, this will increase the accuracy of the segmentation; to do that, the network obtains a fused value from three neural confidences values: 1) Segmentation Point Validation (SPV), 2) Right Character Validation (RCV), and 3) Central Character Validation (CCV) which will assess each PSP separately. The input vectors of the neural network are calculated based on Direction Feature (DF), DF considers much more suitable for Arabic Scripts. AHS and neural-based segmentation techniques have been implemented and tested by local benchmark database.

Keywords:

Arabic handwriting recognition, neural networks, Arabic heuristic segmenter.

1. INTRODUCTION

The concept of handwriting recognition can be divided according to [Pla01a] into two main areas, these areas are on-line and off-line. An off-line Arabic handwriting segmentation and recognition is one of the most challenging researches because there are different variations in handwriting [Naw01a], it is an approach that interprets characters, words and scripts that have been written at common surface (i.e. paper). On the other hand, on-line handwriting recognition refers to automatically recognizing the handwritten characters using real-time information such as pressure and the order of strokes made by a writer usually employing a stylus and pressure sensitive tablet [Cas01a, Lor01a].

The segmentation [Bal01a, Man01a] of Arabic handwritten characters have been an area of great interest in the past few years [Blu00a]. One typical approach in the literature is "over-segmentation" which is known as dissecting the word image based on shape features of the image into a sufficient number of components; so that no merged characters remain [Yan01a, Xia01a]. One of the major problems following over-segmentation is correctly discard the invalid segmentation points and remained the valid points, to determine the valid segmentations, many of researches are studied by merging segments of the image and invoking a classifier to score the combinations, the most techniques employ the optimization algorithms that making use some sort of lexicon-driven and dynamic programming technique [Blu00a]. The best way to evaluate the over-segmentation is use the neural networks [Fan01a], the most common family of neural networks for pattern classification recognition is Feed-Forward Back-Propagation network (FFBP) which is very simple and effective to implement, it has been applied successfully to different applications domains, such as pattern recognition, controlling, prediction, system identification, etc. [Bil01a], the weight inputs transmits to the neurons in the first layer and the neurons transmits their outputs to the neurons in to the next layer, etc., the network not contain any cycles or loop as an advantage [Abd01a].

2. LITERATURE REVIEW

Earlier art showed segmentation of both machine-print and handwriting. In 1980, Nouh *et al*. suggested a standard Arabic character set to facilitate computer processing [Nou01a]. Sami El-Dabi *et al.* used segmented characters based on invariant moments only after they were recognized. Recognition was attempted on regions of increasing width until a match was found [Dab01a]. Yamin and Aoki presented a two-step segmentation system which used vertical projection onto a horizontal line followed by feature extraction and measurements of character width [Ymi01a]. Al-Badr and Haralick presented a holistic recognition system based on shape primitives that were detected with mathematical morphology operations [Bad01a]. Hamami and Berkani developed a structural approach to handle many fonts, and it included rules to prevent over-segmentation [Ham00e]. Al-Qahtani and Khorsheed presented a system based on the portable Hidden Markov Model Toolkit [Qah01a]. Srihari and Ball, applied heuristic techniques for image processing representation of the binary image counter and removal of noise and dots [Sri01a] Hamad and Zitar [Ham00c] applied new fusion equations in order to enhance the segmentations processes. Hamad [Ham00d] developed a technique that aim to assign the prospective segmentation points which is obtained based on the shape features of word image. On the other hand, many researches are using the feed-forward back-propagation neural network, the origin of this type is used by Rumelhart [Rum01a] in 1986, the application area network of back-propagation algorithm are gained recognition and utilized multiple layers of weight-sum units of the type f = g(w'x + b). Training was done by a form of stochastic gradient descent.

3. PROBLEMS OF ARABIC SCRIPTS

Many researches have been published in the area of handwritten Arabic scripts recognition [Ham00a, Ham00b], so far, the researches haven't been reached to good result because it is considerably harder due to a number of reasons: 1) Arabic is written cursively, i.e., more than one character can be written connected to each other. 2) Arabic uses many types of external objects, such as dots, "Hamza", "Madda", and diacritic objects. These make the task of line separation and segmentation scripts more difficult. 3) Arabic characters can have more than one shape according to their position: initial, middle, final, or stand alone. 4) Characters that do not touch each other but occupy a shared horizontal space that increases the difficulty of segmentation, 5) Arabic uses many ligatures, especially in handwritten text, this makes the segmentation of Arabic scripts even more difficult [Ham00c].

4. SEGMENTATION TECHNIQUE

Arabic Heuristic Segmenter (AHS) or over-segmentation technique aims to assign correct PSP points in the word image [Nic01a]. Following this, a neural confidence-based module has been used to validate these points by obtaining a fused value from three neural confidence values based on Segment Point (SP), Right Character (RC), and Central Character (CC) [Che00a]. Segmentation technique has two advantages; first, reducing the number of missed or bad points, and second, increasing the accuracy of the recognition rate. Since number of segmentation points is optimized by using this technique, the overall accuracy will increase and processing time will reduce [Che00b]. Missed points occur when no segmentation point is determined between two successive characters; besides, bad points refer to the points that could not be used to extract the characters precisely. AHS which was proposed by Hamad [Ham01c, Ham00d] removed the punctuation marks (dots) that hinder identify the correct segment points, this technique helps to detect

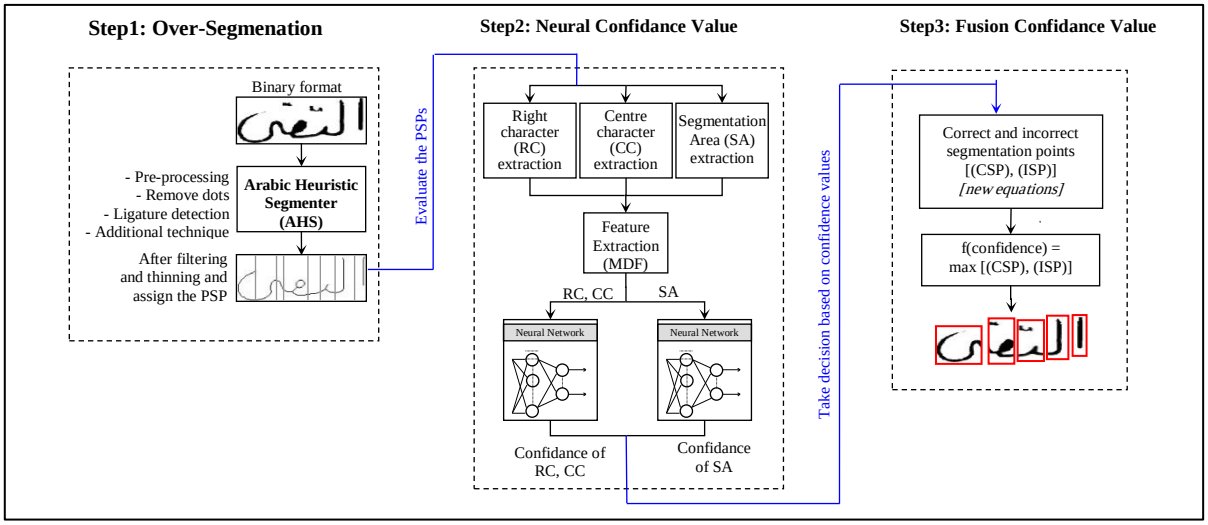


Figure 1. Overview of the segmentation technique

 ISBN 978-80-86943-75-6

the ligatures that connect between two successive characters to obtain the correct segmentation points. Additional techniques such as average character width are applied as well to enhance the results. One of the major problems following over-segmentation is correctly discard the invalid segmentation points and retained the valid one by using neural network, the input vector of the network is extracted based on Modified Direction Feature (MDF) [Blu00b]. Figure 1 illustrates an overview of the entire neural-based segmentation technique.

4.1. Over-Segmentation

Over-Segmentation or AHS employs three techniques: 1) Pre-processing, filtering the word image, and removing the punctuation marks or any redundant components. 2) Ligature detection, a ligature is a small point (stroke) that is used to connect between two characters; the aim of ligature detection is locate these strokes within the "middle region" of handwritten words. 3) Calculates the average character width, the technique aims to add any missing segment points and remove the bad one, an addition technique is detect the close and open holes which is aims to remove any bad points across these holes that considered complete characters. The results of these techniques are word image contains a sufficient number of PSPs; these points will be evaluated by the neural networks later.

4.2. Modified Direction Features (MDF)

Arabic handwritten has a special characteristics such as rotations, curves, and circuits shapes; so, the suitable features input in the vectors of neural network is direction features, MDF extracts the direction information (feature) from structure of the character contours that determined in each character image, the technique categorizes into four parts: 1) Vertical lines, 2) Horizontal lines, 3) Right diagonal, and finally 4) Left diagonal. This principle is extended so that integrates the direction feature with the technique for calculate the transition features between background pixels (white pixels) and foreground pixels (black pixels). In MDF, Location of Transitions (LTs), and Direction Transition (DT) are calculated at a particular location, therefore, for each transition, a pair of values such as [LT, DT] are stored; this work demonstrated the superiority of MDF for describing the Arabic patterns according to their contour or boundary. More details have been described in [Blu00a, Ham00c].

4.3. Neural-based Validation

As a result of above and after completing the over-segmentation, the post-processing is employed to exclude the bad segment point and remain the correct. The classifier chosen for the validation is a feed-forward neural network trained with the back-propagation algorithm. For experimental purposes, the architectures were modified varying the number of inputs, outputs and hidden units. Three vectors are extracted from the word image to validate each PSP and determine whether correct or note, where the classifier will calculate and output the confidence value for each point, the values represent each of the segmentation area (SA), right character (RC), and center character (CC) and validate all of them based on maximum of confidence value. Therefore, it is possible to validate prospective segmentation points, rather than giving a binary result (valid or invalid) decision whether a segmentation point should be set in a particular region, confidence values are assigned to each segmentation points that are located through feature detection. The confidence value of any segment area should be in the range of 0 and 1.

4.4. Fusion Confidence Values

Fusion confidence value is a set of equations take the final decision (valid or invalid), where are calculated on the basis of the output confidence value of the neural network. New fusion equations are proposed, the extracted areas of these equations are analyzed and described as: Rule 1: Following RC extraction and neural verification, the area is analyzed into two options: 1) If the area is identified by the neural expert as one of 62 possible characters, then the segmentation point is more likely to be a correct segmentation point. 2) If the area is identified as a non-character (rejected), then the segmentation point is more likely to be an incorrect segmentation point. Rule 2: Following CC extraction and neural verification, the area is analyzed into two options: 1) If the area is identified by the neural expert as one of 62 possible characters, then the segmentation point is more likely to be an incorrect segmentation point. 2) If the area is identified as a non-character then the segmentation point is more likely to be a correct segmentation point. Rule 3: Following SA extraction, the area is analyzed into two options: 1) If the neural expert provides a confidence >=0.5, then the segmentation point is more likely to be a correct segmentation point. 2) If the neural expert provides a confidence <0.5, then the segmentation point is more likely to be an incorrect segmentation point

Two possibilities for each fusion are applied, first, calculate Correct Segmentation Point (CSP) where Segmentation Point Validation (SPV) >=0.5 as shown in equation 1; second, calculate Incorrect Segmentation Point (ISP) where SPV<0.5 as shown in equation 2; finally, calculate outcome of the fusion decision based on maximum value between the CSP and ISP as shown in equation 3. If the CSP confidence is greater, then the SP will be set as being correct. Conversely, if the ISP confidence prevails as being larger, the SP will be discarded and no longer used in further processing.

 ISBN 978-80-86943-75-6

1) Correct Segmentation Point (CSP):

if $f_{SPV\_ver}(ft1) >= 0.5$ AND $f_{RCC\_ver}(ft2)$ is a high character confidence AND $f_{CCC\_ver}(ft3)$ is a high non-character confidence, then:

$$f_{CSP}(ft1, ft2, ft3) = f_{SPV\_ver}(ft1) + f_{RCC\_ver}(ft2) + (1-f_{CCC\_ver}(ft3)) \quad (1)$$

2) Incorrect Segmentation Point (ISP):

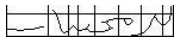
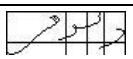
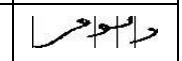
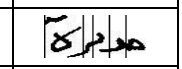
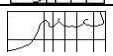
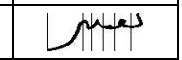
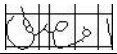
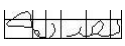
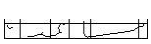
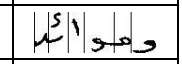
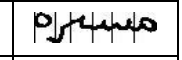
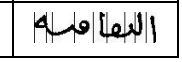
if $f_{SPV\_ver}(ft1) < 0.5$ AND $f_{RCC\_ver}(ft2)$ is a high non-character confidence AND $f_{CCC\_ver}(ft3)$ is a high character confidence, then

$$f_{ISP}(ft1, ft2, ft3) = f_{SPV\_ver}(ft1) + (1-f_{RCC\_ver}(ft2)) + f_{CCC\_ver}(ft3) \quad (2)$$

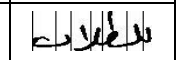
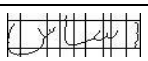
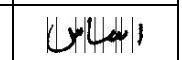
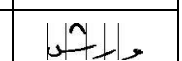
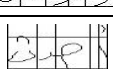
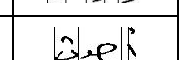
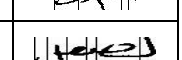
3) Finally, the outcome of the fusion is decided by the following equation:

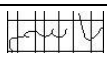
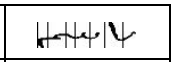
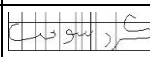
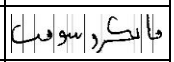
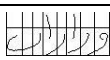
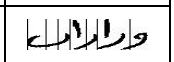
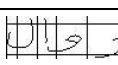
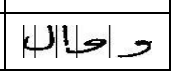
$$f(confidence) = \max [(CSP), (ISP)] \quad (3)$$

Where, $f_{SPV\_ver}(features)$ is confidence value of Segmentation Point Validation, $f_{RCC\_ver}(features)$ is a confidence value for right character, and $f_{CCC\_ver}(features)$ is confidence value for center character (reject neuron output).

Original Word	Over-segmentation	Segmentation
بدمجيات		
وتوفر		
مديرة		
تغيير		
افضل		
تقدية		
مساعدة		
وفوائد		
مسيرة		
الثقافية		

(a) successful segmentation

Original Word	Over-segmentation	Segmentation
للطلاب		
أساس		
ورش		
أهدف		
لصقل		

باسم		
فايكرو سوفت		
وزارات		
العديد		
وحال		

(b) unsuccessful segmentation

Figure 2. Segmented sample of word images

5. EXPERIMENTAL RESULTS

The experiments here used the neural confidence-based module for validating the PSPs which are obtained from AHS (over-segmentation). Segmentation performance is measured based on three types of segmentation errors: "over-segmentation", "missed" and "bad" metrics. Over-segmentation refers to a character that has been divided into more than three components. A "missed" error occurs when no segmentation point is found between two successive characters. The "bad" error refers to a segmentation point that could not be used to extract a character precisely.

5.1. Handwriting Database

The training and testing patterns samples were obtained and extracted from twenty different persons, all words are selected randomly. They were asked to write down two paragraphs contains all status of Arabic characters. These paragraphs scanned at 200 pixels per inch. The size of training set was 620 characters (10 writers x 62 characters), and size of testing set was 425 words, more details about the database see www.acdar.org.

5.2. AHS Segmentation Performance

The total numbers of segmentation points in the 425 testing word samples are 3080. Table 1 shows the segmentation performance of the AHS technique, see [Ham00d] for more details about this results.

Result		Segmentation Error Rates			
		Over Seg.	Missed	Bad	Bad/ overlap
Totals		29	18	552	*26*
%		0.94%	0.58%	17.92%	*0.84%*
With overlap	Total	599			
	%	19.45%			
Without overlap	Total	573			
	%	**18.60%**			

Table 1. AHS segmentation error

5.3. Neural-based Performance

Results of the neural-based segmentation technique were calculated based on the number of correct and incorrect identified of segment point in word samples. Neural network verifies whether

 ISBN 978-80-86943-75-6

segmentation points are valid or invalid based on neural confidence-based module. If the network output a height confidence value this indicated that a point is a valid segmentation point; a low confidence value indicated that a point should be ignored, Table 2 illustrates the overall results of the technique.

Result		Correctly Identified		Incorrectly Identified		
		Valid	Invalid	Valid	Invalid	Invalid overlap
Totals		2011	729	192	148	*40*
%		65.29%	23.67%	6.23%	4.81%	*1.30%*
With overlap	Total	2740		340		
	%	**88.96%**		**11.04%**		
Without overlap	Total	2780		300		
	%	90.26%		9.74%		

Table 2. Results of neural-based segmentation technique

The above results describe the recognition rate for the neural networks. To enhance these rates, the number in the testing set must be increased at least two or three-fold, that will help improving overall segmentation accuracy, Figure 3 illustrates the characters recognition rates of the neural-based segmentation technique.

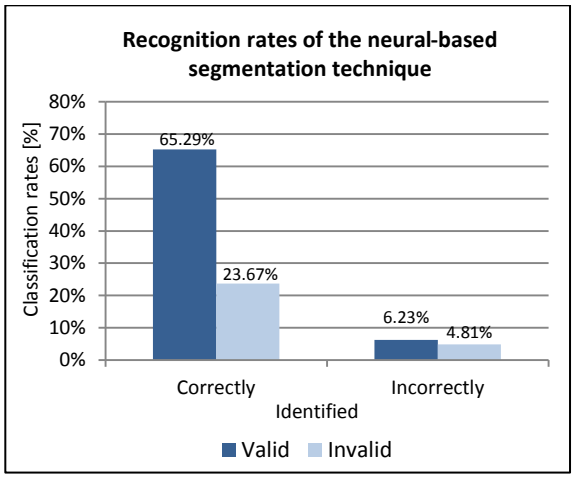


Figure 3. Recognition rates for all different neural networks.

Table 3 shows the summary of the literature results and the comparisons with the paper's results.

Reference	Accuracy	Language / Databases
Blumenstein, Myer [Blu00c]	75.28%	• Cursive English handwriting • CEDAR database
Hamid, Alaa [Ham00a]	69.72%	• Arabic handwriting • Local database: 360 addresses, 4000 images
Cheng, Chun Ki [Che01a]	85.74%	• Cursive English handwriting • CEDAR database: test 1031 from 1718 SP
Khateeb, Jawad [Kha01a]	85.00%	• Arabic handwriting • Local database: 200 images, sub-words SP
Hamad, Husam Al [Ham00d]	82.98%	• Arabic handwriting • Local database: 500 images
This paper	88.96%	• Arabic handwriting • Local database: 425 images

Table 3. Compare the results with the literature

6. CONCLUSIONS

This paper investigates collection of techniques aims to segmenting the Arabic handwritten scripts, new fusion equations, and heuristic technique are developed, the technique splits the word image into a sufficient number of components, in order to separate the word image into its characters, the technique called "over-segmentation" or Arabic heuristic segmenter (AHS). Modified Direction Features (MDF) is also employed which is considered a promised technique for Arabic scripts, MDF extracts the input vector feature of the neural network, the AHS provides better inputs to the subsequent neural validation process. Promised results were obtained in this study may increase the performance of a segmentation-based handwriting recognition systems. In the future, a larger size of training set will investigated in order to improve the results of the classifiers as well as reduce the errors.

7. REFERENCES

[Abd01a] Abdalla, O.A., and Zakaria, M.N., and Sulaiman, S., and Ahmad, and W.F.W.: A comparison of feed-forward back-propagation and radial basis artificial neural networks: A Monte Carlo study, Information Technology (ITSim), vol. 2, pp.994-998, 2010.

[Bad01a] Al-Badr, B., Haralick, R.: A Segmentation-Free Approach to Text Recognition with Application to Arabic Text, International Journal on Document Analysis and Recognition, vol. 1, pp. 147–166, 1998.

[Bal01a] Ball, G., Srihari, S., Srinivasan, H.: Segmentation-Based and Segmentation-Free Methods for Spotting Handwritten Arabic Words, In: IWFHR, 2006.

[Bil01a] Bilski, J.: The Ud Rls Algorithm for Training Feedforward Neural Networks, Int. 1. Appl. Math. Comput. Sci., pp. 115-123, 2005.

[Blu00a] Blumenstein M., Liu X.Y., Verma, B.: An investigation of the modified direction feature for cursive character recognition. Pattern Recognition. vol. 40(2), pp. 376-388, 2007.

[Blu00b] Blumenstein, M., Liu, X.Y., Verma, B.: A Modified Direction Feature for Cursive Character Recognition. International Joint Conference on Neural Networks. Budapest, Hungary, pp. 2983-2987, 2004.

[Blu00c] Blumenstein, Myer. Intelligent Techniques for Handwriting Recognition, School of

Information Technology, PhD Dissertation, Griffith University-Gold Coast Campus, Australia, 2000.

[Cas01a] Casey, R., Lecolinet, E.: A survey of methods and strategies in character segmentation. IEEE Trans. Pattern Analysis and Mach. vol. 18, pp. 690-706, 1996.

[Che00a] Cheng, C.K., Blumenstein, M.: The Neural-based Segmentation of Cursive Words using Enhanced Heuristics. In: Eighth International Conference on Document Analysis and Recognition, pp. 650-654, 2005.

[Che00b] Cheng, C.K., Liu, X.Y., Blumenstein, M., Muthukkumarasamy, V.: Enhancing Neural Confidence-Based Segmentation for Cursive Handwriting Recognition. In: 5th International Conference on Simulated Evolution and Learning Busan, Korea, SWA-8, CD-ROM Proceedings, 2004.

[Dab01a] El-Dabi, S., Ramsis, R., Kamel, A.: Arabic Character Recognition System: A Statistical Approach for Recognizing Cursive Typewritten Text, Pattern Recognition, vol. 23, pp. 485–495, 1990.

[Fan01a] Fan, X., Verma, B.: Segmentation vs. Non-Segmentation Based Neural Techniques for Cursive Word Recognition. An Experimental Analysis. International Journal of Computational Intelligence and App. vol. 2(4), p.p. 377-384, 2002.

[Ham00a] Hamid, A., Haraty, R.: A Neuro-Heuristic Approach for Segmenting Handwritten Arabic Text, In: ACS/IEEE International Conference on Computer Systems and Applications, p.p. 0110, 2001.

[Ham00b] Hamid, A., Haraty, R.: Segmenting Handwritten Arabic Text. ACIS International Journal of Computer and Information Science, vol. 3 (4), 2002.

[Ham00c] Hamad, H.A., Zitar, R.: Development of an efficient neural-based segmentation technique for Arabic handwriting recognition. Pattern Recognition Journal. ELSEVIER. vol. 43, Issue 8, p.p. 2773-2798, 2010.

[Ham00d] Hamad, Husam A. Al: Over-segmentation of handwriting Arabic scripts using an efficient heuristic technique, In: Wavelet Analysis and Pattern Recognition (ICWAPR), IEEE, pp.180-185, 2012.

[Ham00e] Hamami, L., Berkani, D.: Recognition System for Printed Multi-font and Multisize Arabic Characters, the Arabian Journal for Science and Engineering, vol. 27, pp. 57–72, 2002.

[Kha01a] Jawad H AlKhateeb, and Jianmin Jiang, and Jinchang Ren, and Stan S Ipson. Component-based Segmentation of Words from Handwritten Arabic Text, Proceedings of World Academy of Science, Engineering and Technology, ISSN, vol. 31, pp. 1307–6884, 2008.

[Lor01a] Lorigo, L., Govindaraju, V.: Off-line Arabic Handwriting Recognition: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28 (5), p.p. 712–724, 2006.

[Man01a] Mansour, M., Benkhadda, M.: Optimized segmentation techniques for Arabic handwritten numeral character recognition. In: SITIS, p.p. 96-101, 2005.

[Naw01a] Nawaz, S.N., Sarfraz, M., Zidouri, A.; Al-Khatib, W.G.: An approach to offline Arabic character recognition using neural networks. Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003 10th IEEE International Conference on, vol. 3, p.p. 1328-1331, 2003.

[Nic01a] Nicchiotti, G., Scagliola, C.: A Simple and Effective Cursive Word Segmentation Method. Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, Amsterdam, pp. 499-504, 2000.

[Nouh01a] Nouh, A., Sultan, A., and Tolba, R.: An Approach for Arabic Characters Recognition, *J.Eng. Sci., Univ. Riyadh*, vol. 6, pp. 185–191, 1980.

[Pla01a] Plamondon, R., Srihari, S.N.: On-Line and Off-Line Handwriting Recognition. A Comprehensive Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, p.p. 6384, 2000.

[Qah01a] Al-Qahtani, S., Khorsheed, M., A HTK-Based System to Recognise Arabic Script, in Proc. 4th IASTED International Conference on Visualization, Imaging, and Image Processing. Marbella, Spain: ACTA Press, 2004.

[Rum01a] Rumelhart, David, E., Hinton, Geoffrey, E.: Williams, Ronald J., "Learning representations by back-propagating errors", Nature, vol. 323(6088), pp. 533–536, 1986.

[Sri01a] Srihari, S., Ball, G.: An Assessment of Arabic Handwriting Recognition Technology, in IWFHR, CEDAR Technical Report TR–03–07, 2007.

[Xia01a] Xiao, X., Leedham, G.: Knowledge-based Cursive Script Segmentation. Pattern Recognition Letters, vol. 21, pp. 945-954, 2000.

[Yan01a] Yanikoglu, B., Sandon, P.A.: Segmentation of Off-Line Cursive Handwriting using Linear Programming. Pattern Recognition, vol. 31, pp. 1825-1833, 1998.

[Ymi01a] Ymin, A., Aoki, Y., On the Segmentation of Multi-font Printed Uygur Scripts, in Proc. 13th International Conference on Pattern Recognition, vol. 3, pp. 215–219, 1996.