

Component-based Segmentation of Words from Handwritten Arabic Text

Jawad H AlKhateeb, Jianmin Jiang, Jinchang Ren, and Stan S Ipson

Abstract—Efficient preprocessing is very essential for automatic recognition of handwritten documents. In this paper, techniques on segmenting words in handwritten Arabic text are presented. Firstly, connected components (ccs) are extracted, and distances among different components are analyzed. The statistical distribution of this distance is then obtained to determine an optimal threshold for words segmentation. Meanwhile, an improved projection based method is also employed for baseline detection. The proposed method has been successfully tested on IFN/ENIT database consisting of 26459 Arabic words handwritten by 411 different writers, and the results were promising and very encouraging in more accurate detection of the baseline and segmentation of words for further recognition.

Keywords—Arabic OCR, off-line recognition, Baseline estimation, Word segmentation.

I. INTRODUCTION

HANDWRITTEN recognition plays essential roles in many applications, such as office automation, cheque verification, mail sorting, and a large variety of banking, business as well as natural human-computer interaction.

In general, this task can be divided into on-line based or off-line based systems. For on-line applications, the computer can trace the process of writing, hence the strength and sequential order of each segment when it is written can be recorded for recognition. While in offline applications, only a digital image is available hence it is more difficult. In this paper, we emphasize on offline recognition of handwritten Arabic text.

Arabic is written by more than 250 million people [1],[2],[3]. Unlike many other languages such as Latin, Chinese, and Japanese scripts which have been widely investigated, recognition of handwritten Arabic text remain a challenge task as there is very limited work reported. By nature, Arabic script is cursive, which makes its recognition rate lower than that of Printed Latin. Arabic text is written from right to left, and it has 28 basic letters in which 16 of them have dots. The dots can be one dot, two dots, or three dots. The dot(s) can be below or above the baseline and accordingly form different

semantics. Therefore, detection of the baseline is one of the most important steps for Arabic text recognition

Basically, there are two different categories of systems for the recognition of Arabic scripts, i.e. segmentation based and segmentation free based. In the first category, words need to be further segmented into characters or letters and these characters are then used for recognition, this is known as analytical approach. While, the second category does not need segmentation and the word images are taken as a whole on recognition, this is known as global approach where the recognition is globally performed. The global approach makes the recognition process simpler by avoiding the difficulty in character segmentation.

The accuracy of a recognition system depends on the quality of input images and effective preprocessing. Once the sample image is acquired, pre-processed is required to enhance the signal for better performance. Pre-processing usually includes many relevant techniques like thresholding, skew/slant correction, noise removal, thinning, baseline estimation and segmentation of words. Except for baseline detection, we also focus on words segmentation.

The work for Arabic script recognition has started more than two decades ago. Almuallim and Yamaguchi [4] proposed a structural recognition technique for Arabic handwritten words which were segmented into strokes. The strokes were classified and combined into characters according to their features. However, their system showed a failure in most cases due to incorrect segmentation of words. Amin and Alsadoun [5] proposed techniques using binary tree to segment printed Arabic text into characters. Recognition of hand printed Arabic characters are introduced in [6] and [7]. Abuhaiba [8] dealt with some problems in the processing of binary images of handwritten text documents, such as extracting lines from pages, which is found to be powerful and suitable for variable handwriting. Abuhaiba et al. [9] introduced a novel offline cursive Arabic script recognition system to recognize offline handwritten cursive script having high variability based on segmentation based system. In their system, a single component strokes were extracted. Khorsheed M S [10] presented a new method on off-line recognition of handwritten Arabic script. The method does not require segmentation into characters, and is applied to cursive Arabic script. The method decomposed the skeleton of the word into an observation sequence. The method trains a single hidden Markov model (HMM) with structural features. HMM is also used in Alma'adeed et al [11] for unconstrained Arabic handwritten word recognition. In Motawa et al [18], mathematical morphology is applied to segment Arabic words

Jawad AlKhateeb is a PhD student at the Department of Electronic Imaging and Media Communications, School of Informatics, University of Bradford, Bradford, UK (e-mail: j.h.y.alkhateeb@bradford.ac.uk).

Jianmin Jiang is a professor of Digital Media at the School of Informatics, University of Bradford, Bradford UK (e-mail: j.jiang1@bradford.ac.uk).

Jinchang Ren is a research assistant at the School of Informatics, University of Bradford, Bradford, U.K. (e-mail: j.ren@bradford.ac.uk).

S S Ipson is a senior lecturer the School of Informatics, University of Bradford, Bradford UK (e-mail: s.s.ipson@bradford.ac.uk).

into its characters which is based on the assumptions that characters are usually connected by horizontal lines. Lorigo and Govindaraju [19] presented a character segmentation system which used derivative information in a region around the baseline to over-segment the words.

II. IFN/ENIT DATABASE

Any recognition system needs a large database to train and test the system. Real data from banks or the post code are confidential and inaccessible for non commercial research. Although some work was conducted in Arabic handwritten words, but generally they had small databases of their own or the presented results on databases which were unavailable to the public. Consequently, there was no benchmark to compare the results obtained by researches. The IFN/ENIT database [13], available for free, is very important in this context as it has been used as a standard test set in such a context [12].

The IFN/ENIT was published by the Institute of Communication Technology (IFN) at Technical University Braunschweig in Germany and the Ecole National d'Ingenieurs de Tunis (ENIT) in Tunisia. The database consists of 946 Tunisian town/villages names together with their postcode. In total 411 people were selected as writers to put their names. Also each writer was asked to fill a form with handwritten pre-selected names of Tunisian town/villages with the corresponding postcode. All the forms were scanned with 300dpi and converted to binary images. The images are divided into four sets so that researches can use some of them for training or testing, respectively..

III. SEGMENTATION THE WORDS

A. Baseline Detection

Before segmenting the Arabic words, we need detect the baseline as it is believed that this baseline is very essential in analyzing Arabic text [1-2]. Since the Arabic letters are usually written along the baseline, hopefully there should be a peak is the baseline position when we project the written line along the vertical axis of the image [19]. In most of the cases this assumption is true. However, for some special cases, it goes wrong as shown in Fig. 1(a) in which the detected baseline appears too high. In our improved method [20], the baseline is not decided as the peak in its vertical projection. Instead, we ask its location below the middle line of the image, and then check the peak to decide the baseline. The improved result is shown in Fig. 1(b).

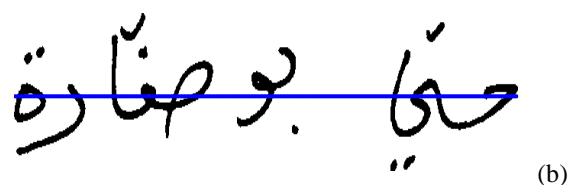
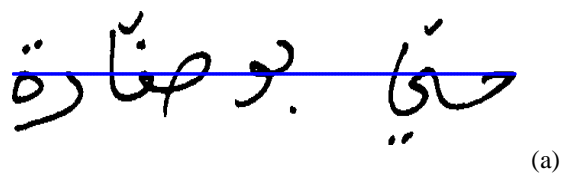

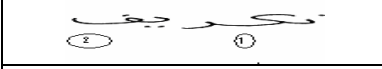
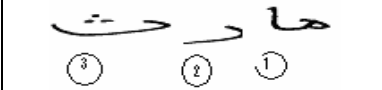


Fig. 1 Baseline detection: original result using only vertical projection (a), and (b) using both vertical projection and knowledge-supporto

B. Extracting Connected Components and Sub-Words

Segmentation is an essential step which separates the text image objects for recognition phase. The typical segmentation for printed binary document is based on the histogram projection analysis, and regrouping the connected components [1-2]. Arabic writing is cursive and is such that words are separated by spaces. However, a word may contain several sub-words which are a portion of the word including one or more connected letters. Table I shows three Arabic words consists of one, two, and three subwords respectively.

TABLE I
ARABIC WORDS WITH SUBWORDS

Image	Number of Sub-words
	1
	2
	3

The connected components (ccs) for the line image must be determined. The ccs are rectangular boxes bounding together regions of connected objects. The objective of the ccs phase is to form rectangles around the connected object on the image. The algorithm used to obtain the ccs are the iterative procedure which compares any black pixels in any pair of the line are connected together. Bounding rectangles are extended to enclose any grouping of connected black pixels. Fig. 2(a) shows the output of the ccs.

With extracted connect components, sub-words are segmented as follows. Firstly, small parts like dots in the image are temporally ignored. Secondly, components whose coordinates are overlapped in x-axis are merged to obtain a combined large component, namely sub-word. Thirdly, the distance of each pair of consecutive sub-words is obtained, which will be used to segment words in the next section.

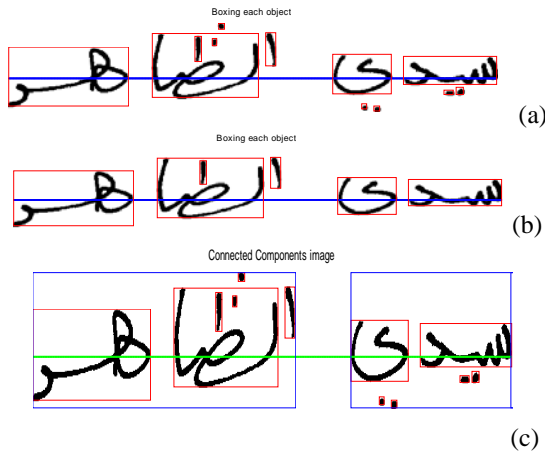


Fig. 2 Examples of extracted connected components (a), sub-words of combined components (b), and detected words (c)

C. Word Segmentation

Arabic writing is cursive; therefore, words and subwords are separated by spaces, so word boundaries are always represented by a space. According to this, distances between each pair of consecutive sub-words are obtained. Normally the distances between words are larger than the distances between subwords, thus words can be segmented by comparing this distance against a suitable threshold.

To determine such a threshold, Bayesian criteria of minimum classification error is employed as follows. Given a distance of d , the probability that represents a separation of word or sub-word is denoted as $p_w(d)$ and $p_{s-w}(d)$, respectively. These two conditional probabilities are obtained by manually analyzing over 200 images containing more than 250 words. Take $p_w(d)$ for example, we find all possible distances to separate a word and then calculate their histogram, and $p_w(d)$ is estimated by this histogram. Certainly, these distances are found on the basis of our obtained vertical projection of the image. Illustrations of both $p_w(d)$ and $p_{s-w}(d)$ are given in Fig. 3.

Afterwards, an optimal distance d_0 is obtained under the Bayesian minimum classification error criteria:

$$d_0 = \arg \min_d (err(d)) \quad (1)$$

$$err(d) = \int_d^\infty p_{s-w}(x) dx + \int_0^d p_w(x) dx \quad (2)$$

Finally, segmentation of words is completed by simply comparing the distance d with this optimal distance or threshold d_0 . If we have $d > d_0$, it refers to two words. Otherwise, it is two sub-words. According to the original image in Fig. 2(a), the final segmented words are shown in Fig. 2(c).

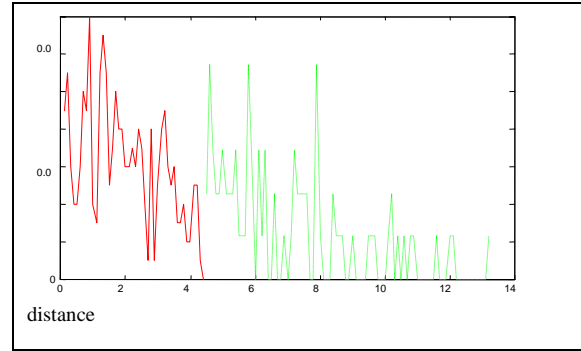


Fig. 3 Illustrations of $p_w(d)$ (in dot) and $p_{s-w}(d)$ (in solid line)

In general, Segmentation points occur where one word ends and another begins. Segmentation is the process of detecting these points and is important or recognition

IV. EXPERIMENTAL RESULTS

This paper is concerned in word segmentation using vertical histogram and connected component analysis. Here, the distances between subwords were measured and compared to an optimal threshold to determine if the distance corresponds to separation of two words or not.

In general, there are different type of errors occur during the process of segmentation either the analytical approach or our approach. These errors can be summarized as:

- 1) Over segmentation when the number of segments is greater than the actual number.
- 2) Under segmentation when the number of segments is less than the actual number.
- 3) Misplaced Segmentation when the number of segments is right but the limits are wrong.

We have tested our techniques on 200 images in the test set and the results are summarized in Table II below. Meanwhile, some more results are also presented in Fig. 4.

TABLE II
OVERALL SEGMENTATION RESULTS

No. of Images	Correct Seg.	Under seg.	Over seg.	Misplaced Seg.
200	85%	9%	4%	2%

From Table II we can see the correct segmentation rate for images is 85%, but the segmentation error is 15% which is due to the variation in handwriting, especially irregular spaces between sub-words and words, such as too large spaces between sub-words (which may be wrongly taken as two words and lead to over-segmentation) or too small spaces between words (which will lead under segmentation by incorrectly merging two words together). Examples of these errors are illustrated in Fig. 5. To overcome such errors, additional information like knowledge of the Arabic language is needed which may indicate that certain cases of under-segmentation or over-segmentation is not reasonable. This will be investigated further upon.

	Original Image	Word Segmentation
a		
b		

Fig. 4 Word segmentation result

	Original Image	Word Segmentation
a		
b		
c		

Fig. 5 Failure word segmentation result."

V. CONCLUSION

A component-based method is introduced to segment words from handwritten Arabic texts. Since many people have emphasized on either segment-free based method or letters or strokes based approaches, words segmentation has not be well addressed. Here, our work provides a practical way in accurately segmenting words from the text. This is useful and more flexible than segment-free based approaches as it can make good use the common part of images in further recognition. Also, this is simpler and more robust than letter-based methods because the latter has much difficulty in effective segmenting of arbitrary handwritten characters. We have found that distance information is very essential in segmenting words, yet some improvements are still desired in considering knowledge of the language. Further investigations include applying pre-recognition into fine segmentation and probability-based reasoning in recovering segmentation errors in a recognition stage.

REFERENCES

- [1] A. Amin. "Offline Arabic character recognition: The state of the art". *Pattern Recognition*, vol. 3, pp. 517-530, 1998.
- [2] L. M. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: a survey", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 712-724, 2006.
- [3] M.S. Khorsheed, "Off-Line Arabic Character Recognition – A Review", *Pattern Analysis & Applications*, vol.5, pp. 31-45, 2002.
- [4] H. Al-Muallim and S Yamaguchi. "A method of recognition of Arabic cursive handwriting". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, pp. 715-722, 1987.
- [5] A. Amin and H. Alsadoun, "A new segmentation technique of Arabic text.", *IEEE Trans. Pattern Recognition*, Vol.2, pp. 441-445, 1992.

- [6] A. Amin and H. Alsadoun, "Hand printed Arabic Character Recognition System", *IEEE Trans. Pattern Recognition*, Vol. 2, pp.536-539, 1994.
- [7] I. S. I. Abuhaiba and P. Ahmed, "Restoration of temporal information in off-line arabic handwriting," *Pattern Recognition*, vol. 26, pp. 1009-1017, 1993.
- [8] I. S. I. Abuhaiba, M. J. J. Holt, and S. Datta, "Processing of binary images of handwritten text documents," *Pattern Recognition*, vol. 29, pp. 1161-1177, 1996.
- [9] I. S. I. Abuhaiba, M. J. J. Holt, and S. Datta, "Recognition of Off-Line Cursive Handwriting," *Computer Vision and Image Understanding*, vol. 71, pp. 19-38, 1998.
- [10] M. Khorsheed, "Recognising handwritten Arabic manuscripts using a single hidden Markov model", *Pattern Recognition Letters*, vol. 24, pp. 2235-2242, 2003.
- [11] S. Alma'adeed, C. Higgins, and D. Elliman, "Off-line recognition of handwritten Arabic words using multiple hidden Markov models", *Knowledge-Based Systems*, vol. 17, pp. 75-79, 2004.
- [12] F. Farooq, V. Govindaraju, and M. Perrone, "Pre-processing Methods for Handwritten Arabic Documents", *proc. Int'l conf. Document Analysis and Recognition*, vol. 1, pp. 267-271, 2005.
- [13] IFN/ENIT - Database of Arabic Handwritten words, Institute of Communications Technology, Technical University Braunschweig, Germany.
- [14] M. Pechwitz, and V. Margner. "Baseline Estimation for Arabic Handwritten Words". *International Workshop on Frontiers in Handwriting Recognition*, pages 479-484, 2002.
- [15] H. Al-Rashaideh, "Preprocessing phase for Arabic Word Handwritten Recognition", *Information Transmissions in Computer Networks*, vol.6, pp. 11-19, 2006.
- [16] M.Syiam, T.M. Nazmy, A.E. Fahmy, H. Fathi, and H. Ali, "Histogram Clustering and Hybrid Classifier for Handwritten Arabic Characters Recognition", *Proc. IASTED Int. Multi-conf. Signal Proc., Pattern Recognition and Applications*, pp 44-49, 2006.
- [17] B. Al_Badr, and R. Haralick, "Segmentation-Free Word Recognition with Application to Arabic", *proc. Int'l conf. Document Analysis and Recognition*, vol. 1, pp. 355-359, 1995..
- [18] D. Motawa, A.Amin, and R. Sabourin, "Segmentation of Arabic Cursive Script", In *Proceeding of the 4th International conference Document Analysis and Recognition*, vol. 2, pp. 625-628, 1997.
- [19] L. Lorigo and V. Govindaraju, "Segmentation and pre-recognition of Arabic handwriting," *proc. Int'l conf. Document Analysis and Recognition*, vol. 2, pp. 605-609, 2005.
- [20] J. AlKhateeb, J. Ren, S. S. Ipson and J. Jiang: "Knowledge-based Baseline Detection and Optimal Thresholding for Words Segmentation in Efficient Pre-processing of Handwritten Arabic Text". *International Conference on Information Technology: New Generations*, pp.1158-1159, 2008.

Jawad AlKhateeb received B.Sc degree in Electrical Engineering from Louisiana Tech University, USA, in 1991, MSc degree in Electrical Engineering with an emphasis on Microprocessors based systems from Louisiana Tech University, USA, in 1992. He was a Lecturer (1993–2006) with the School of Electrical and Computer Engineering at Applied Science University, Jordan, and Ajman university of Science and Technology, UAE. Currently, he is a PhD student at the University of Bradford, UK. His research interest includes image processing, document and character recognition,

Jianmin Jiang received B.Sc degree from Shandong Mining Institute, China, in 1982, M.Sc degree from China University of Mining and Technology in 1984, and PhD from the University of Nottingham, UK, in 1994. From 1985 to 1989, he was a lecturer at Jiangxi University of Technology, China. In 1989, he joined Loughborough University, UK, as a visiting scholar and later moved to the University of Nottingham as a research assistant. In 1992, he was appointed a lecturer of electronics at Bolton Institute, UK, and moved back to Loughborough University in 1995 as a lecturer of computer science. In 1997, he was appointed as a full professor at the School of Computing, University of Glamorgan, Pontypridd, UK. He joined University of Bradford in 2002 as a professor of Digital Media at the School of Informatics, University of Bradford, UK. In 2004, he was appointed as an adjunct professor at the Southwest China University, Chongqing, China. He is a fellow of IEE and fellow of RSA in the UK. His research interests include visual information retrieval, image/video processing, visual content

management, Internet video coding, stereo image coding and neural network applications. He has published more than 180 refereed research papers.

Jinchang Ren received the B.S. degree in computer software, the M.S. degree in image processing and pattern recognition, and the Ph.D. degree in computer vision from Northwestern Polytechnic University (NWPU), Xi'an, China, in 1992, 1997, and 2000, respectively. He was a Lecturer (1997–2000) and then Associate Professor with the School of Computers, NWPU. From 1998 to 2006, he had held several research positions with the Hong Kong Polytechnic University (Center of Multimedia Signal processing), University of Abertay Dundee (Center for Computer Games and Virtual Entertainment), Kingston University (Digital Imaging Research Center), and the University of Surrey (Center for Vision, Speech and Signal Processing). In November 2006, he joined the School of Informatics, University of Bradford, Bradford, U.K. His research interests focus on multimedia signal processing, pattern recognition, and computer vision, such as content-based video storage and retrieval, archive restoration, and visual surveillance.

Stan S Ipson, MInstP, is a Senior Lecturer in the Electronic Imaging and Media Communications department at the University of Bradford. Originally trained as a physicist, with a first class Honours degree in Applied Physics and a PhD in Theoretical Nuclear Physics, his current research expertise includes image processing, pattern recognition and the design of machine vision systems. His imaging research interests have covered a variety of applications including video-rate de-blurring of millimetre wave images, vision based control of industrial machinery, face recognition, recognition of cursive Arabic handwriting, ASIC based motion detection for surveillance using multiple cameras, photogrammetry applied to rural and urban landscapes and the detection of solar features. He has over 70 book and journal publications. His current research interests include solar imaging, space weather, machine learning, and 3D modelling from 2D images.