Recognition-based Segmentation of Arabic Handwriting

Ashraf Elnagar and Rahima Bentrcia

Department of Computer Science, University of Sharjah
P. O. Box 27272, Sharjah, U.A.E.

Abstract. Several segmentation approaches proposed in the past decades for Arabic handwritings suffer from over-segmentation. This problem decomposes a single letter into small strokes. The aim of this work is to handle this problem using Artificial Neural Networks with a set of combination rules to keep the correct strokes (letters) and combine the over-segmented ones to intact letters in a correct way. After word segmentation, the resulting segments are normalized. Then, a set of features was extracted from each segment and passed to Artificial Neural Network to be recognized. Finally, proposed combination rules were applied to unrecognized strokes and to specific recognized letters. The success rate of the experimental results exceeds 95%.

1 Introduction

Automatic recognition of handwritings becomes a mature subject because of the wide studies done in this field. Two different classes of character recognition systems are defined: on-line and off-line systems. On-line systems recognize handwritings that are entered from a tablet or any sensitive device by a digital pen. Off-line systems deal with images of handwritings stored in a database.

In this paper, we focus on Arabic handwritten characters recognition, [1]. Despite the difficulty faced in recognizing Arabic handwritings, which is represented in the quality of the writing (the poorer the writing is the harder to be recognized), this subject is still under research because of its important applications. So many services need to automate human processes for time and accuracy purposes. One application appears in automating the process of reading and recognizing handwritten accounts names and checks amounts in banks. Sorting incoming mails by recognizing the handwritten addresses is another main application needed in posts. One more application can be seen in retrieving the ancient Arabic handwritten manuscripts.

Our recognition approach is embedded within the segmentation system which is proposed in [2]. After word segmentation, the resulting segments are normalized and a set of twenty features is extracted and fed into an artificial neural network. The recognition outputs can range from class 1 to class 46, as clarified in Table 1, where each class represents the location of a single letter in the word or the location of a group of letters that share similar shape characteristics. Some segments may not be recognized because of the poor writing, so they are combined with the following segment and a new set of twenty features is extracted again and passed to the neural

Elnagar A. and Bentrcia R. (2009).
Recognition-based Segmentation of Arabic Handwriting.
In *Proceedings of the 9th International Workshop on Pattern Recognition in Information Systems*, pages 83-92
DOI: 10.5220/0002179400830092

network. This process is repeated until the segment is recognized. A set of combination rules is also applied to specific recognized characters.

The rest of the paper is organized as follows. Section 2 presents related work. Recognition stage is described in Section 3. Section 4 discusses the experimental results, and Conclusion and future direction are included in Section 5.

2 Related Work

Most of the work on handwriting recognition was done on Latin text. This lack in Arabic handwriting recognition systems is highly related to the difficulty of segmenting words into characters because of the cursive nature of Arabic handwriting. Therefore, Arabic recognition methods can be divided into those which first segment the word to be recognized, and those that recognize the whole word.

Earlier surveys discussed both Arabic printed and handwritten texts, [3, 4, 5, and 6]. One of the segmentation based methods for automatic recognition of printed Farsi, Arabic, and Urdu texts was proposed by Parhami and Taraghi in 1981, [7]. In this approach, sub-words were segmented and recognized according to geometrical features such as concavities, loops, and connectivity. No performance results were reported for this algorithm but smaller type fonts may not recognized perfectly.

In 1986, Amin and Masini proposed a system for segmentation and recognition that used horizontal and vertical projections and shape-based primitives [8]. On 100 multi-font words, it achieved a character recognition rate of 85% and a word recognition rate of 95%.

Gillies et. al. constructed a recognition system for Arabic text, [9], where words were over-segmented using two different splitting methods, then the resulting segments are ordered and combined in groups and sent to a trained neural network which recognized whole characters from the combination options. These were passed to a Viterbi search to predict the word. Using a testing set of 138 page images, digitized to 200X200, the system achieved a recognition rate of 93%. However, this rate degraded to 89% when the same set was used with a size of 100X200.

In 2002, Hamami and Berkani developed a structural approach to handle many fonts and it included rules to prevent over-segmentation [10].

Hidden Markov Model (HMM) is used also to recognize words by using words features. In 2001, Dehghan et. Al split words into overlapping vertical segments [11], then, they extracted column features and passed them to HMM.

Al-Qahtani and Khorsheed presented a system based on Hidden Markov Model Toolkit in 2004, [12, 13]. One system did not require the segmentation stage and recognized Arabic scripts using HTK. The second system decomposed the text into line images and divided each line image into smaller overlapped frames. Then it extracted statistical features from each frame and passed them to HTK.

Two segmentation free recognition methods appeared in 1995 by Al-Badr and Haralick. In the first system, [14], the whole word was recognized by detecting a set of shape primitives which matched to a constrained set of symbol models. The recognition rate was 99.7% for synthetically degraded symbols and 94.1% for scanned symbols. For isolated words, the system achieved 99.4% for noise-free words, 95.6% for synthetically degraded words, and 73% for scanned words. The

second system was developed to recognize machine printed Arabic words without prior segmentation. The idea was based on shape primitives that were detected with mathematical morphology operations, [15]. The recognition rate was 99.4% for noise-free texts and 73% for scanned texts.

Khorsheed and Clocksin proposed in 1999 another holistic system where features were extracted from a word's skeleton for recognition without prior segmentation [16].

In 2000, Amin introduced another holistic approach where global features such as loops and peaks were extracted from the input word [17], and passed to the C4.5 machine learning system to generate a decision tree for classifying the word. The success rate of the system was 92% using 1000 Arabic words with different fonts.

Another method was presented by Pechwitz and Maergner [18], where the recognition system was based on a semi-continuous 1-dimensional HMM. From each input word, features were collected using sliding window approach. The recognition results achieved 89%, using the IFN/ENIT database of Arabic handwritten words for testing.

In this work, an effective segmentation method for Arabic handwriting was developed. The method used a multi-agent approach to segment words and relied on recognition to verify the validity of the candidate segmentation points. Comparing the previous methods of segmentation approaches and our approach, this segmentation method is not only resolved the shortcomings of the previous related methods but also achieved better results by avoiding under segmentation. This depended on the high performance of the agents and the right decision to select artificial neural network with combination rules which improved detecting the candidate segmentation points.

3 Segmentation Stage

Our segmentation system, which we proposed in [2], was basically based on a multi-agent approach to identify the segmentation points.

Initially, the image of Arabic handwritten text was binarized and cleaned from noise. Then, the text was segmented into lines and each line was segmented into words. The resulting words were thinned and the main connected components in each word were determined and passed to agents that extracted three types of feature points before starting their work.

The identification of initial cutting points strongly depends on seven agents. Six agents are major, which are: loop agent, letter Seen agent, under-baseline-cavities agent, above-baseline-right-cavity agent, above-baseline-left-cavity agent, and above-baseline-narrow-left-cavity agent. The other agent which is the baseline agent is a minor one since it was used by major agents to facilitate their task. First, the agents detected regions that look like some of Arabic characters, these regions were subtracted from the whole word and the remaining parts were left for further processing. Next, all end points features were extracted from the remaining regions and an initial cutting point was inserted between every two successive end points. Finally, a set of filtering rules was applied to remove the extra segmentation points. The experiments reported very good results where the success rate was 86%.

4 Recognition Stage

This phase is very important in our segmentation system, [2]. Since segmenting words into characters is a challenging task, especially for Arabic handwritings, a verification tool is needed to measure the segmentation performance. Artificial neural network was selected to decide if the resulting segment is a letter or a stroke and then needs further processing.

Generally, artificial neural networks are very common in pattern recognition field. Our decision to use ANN as a recognition model was based on the excellent features that it possesses compared to other recognition tools. A well-trained neural network can perform complex functions and solve challenging problems that are difficult for conventional computers or human beings since it is based on learning what it sees. In addition, neural networks can be modified easily and retrained when the requirements of the problem are changed. Finally, its integration property allows several recognition tools to work properly and cooperate with neural networks. This feature may increase the efficiency of the problem solution.

The following sections describe the main steps in our approach.

4.1 Features Extraction

In this step, each segment image is converted into numerical features which describe the segment. The feature extraction methods used in character segmentation systems are probably the most important factor in achieving a good segmentation/recognition rate. After segmenting the word, its output segments are normalized into 250X250 images. Then, twenty structural features are extracted from them. Fifteen Fourier descriptors are extracted from the segments contour and normalized to remove character variations in shift, size, and rotation, [19, 20]. The other five features include number of loop, number of black points to total number of points ratio, the existence of connection to the right and left of the segment [21], and height o width ratio.

A different number of Fourier descriptors are tested and the final set includes 15 descriptors. The selection of these features was based on their ability of describing the general shape of any closed curve such as characters by a set of Fourier coefficients. Suppose that a character consists of a sequence of points (xi, yi), where i=1, 2, …, N, and N is the number of points in character's boundary. Each of these points can be represented as a complex number: a(n) = x(n)+ i*y(n). The discrete Fourier transformation u(n) represents the coordinate sequence a(n). The first 15 coefficients (descriptors) are selected as our features. This is referred to that the general properties of the character shape are kept in the first (low) coefficients. Because characters varied in size, location and maybe rotation angle, Fourier descriptors can be manipulated to be character rotation, scale, and shift invariant. To make Fourier descriptors rotation and shift invariant, only their absolute values are used, and to make them scale invariant, the coefficient are normalized by dividing them by the first coefficient a(1), [19].

4.2 Reconstruction and Recognition

Artificial neural networks are computational models which take their inspiration from the models and theories of the human brain. The most popular neural network is the multilayer feed-forward network where neurons are grouped as layers and connections between neurons in consecutive layers are permitted. The inputs are fed from the input layer and outputs are at the output layer.

In this work, after images normalization, a vector of 20 features is extracted from each segment image and classified using a feed forward neural network trained by back-propagation learning algorithm [1]. The structure of this ANN, consists of four layers: one input layer of 20 neurons, two hidden layers of 100 neurons and one output layer of 46 neurons. The neurons in the hidden layers and the output layer are working using tan-sigmoid and linear algorithms, respectively, and the network is trained using traincgf function. The final selection of the ANN's structure and the used algorithms was determined after trying so many other structures and testing several algorithms. The ANNs which are trained using 'traincgf' give better results compared to those that use other training algorithms. Moreover, traincgf has smaller storage requirements and faster convergence in some recognition problems.

The 46 outputs represent the classes that each segment may belong to, and each class includes letters that have similar shape (body) in a specific location in the word; in the beginning of the word, in the middle, in the end, or isolated. The list of output classes appears in Table 1.

Table 1. The Output Classes of the Proposed Recognition System.

1	ا	17	سد شد	31	فـ قـ
2	أ	18	س ش	32	ف ق
3	بـ تـ ثـ	19	صـ ضـ	33	ک
4	بـ نـ ثـ	20	صـ ضـ	34	ک
5	ب ت ث	21	ص ض	35	ھ
6	ب ت ث	22	ص ض	36	ـہ
7	خ ح ج	23	ظـ طـ	37	ہ
8	حـ جـ خـ	24	ظ ط	38	ہ
9	خ ح ج	25	ع غ	39	ہ
10	خ ح ج	26	عـ غـ مـ فـ قـ	40	ہ
11	ـل ـر ـل	27	ع غ	41	ہ
12	د ر ل	28	ع غ	42	م
13	ي	29	فـ قـ	43	ں
14	ي	30	ھ	44	ن
15	ش س			45	ۓ
16	لله			46	و

The proposed ANN is trained using 2000 characters; more than 40 characters from each class, written by different people. Then, testing was accomplished by selecting examples from each class and passing them to the ANN. A total of 250 characters were used as testing examples. The obtained recognition rate exceeds 87%.

4.3 Restoration and Combination

This step is required when the word is over-segmented and additional segmentation points were determined. As a result, pseudo characters that passed to the neural network are not correctly recognized. To remedy this situation, the extra segmentation points are removed and the adjacent segments are combined and passed again to the neural network, [22]. This process is repeated until the candidate character is recognized.

A preprocessing step was applied first to remove segmentation points that yield to segments with width less than a threshold. This process eliminates most of the strokes which wrongly found in letters such as 'ain ع ' and 'haa ح'. The following examples depict this case in Fig. 1. As clarified in the figure, the small segment of letter 'ع' in the word 'الشرايع', and of the letter 'و' in the word 'صحراوي', and of the letter 'ء' in the word 'الرجاء' are eliminated and combined to their related segments to form complete letters.

Fig. 1. The word before and after extra segmentation points.

Class no.	Double Segmentation	Triple Segmentation
3		
6		
13		
24		
29		
10		
18		
21		
22		

Fig. 2. The Double and Triple Segmentation of Some Characters.

Because word segmentation is a precise process, a set of rules is used to cooperate with the embedded recognition system in order to keep the correct segments (characters) and combine the wrong ones in a correct way.

The combination rules are based on the recognition results of segments. As observed in the segmentation stage, a letter is segmented in the worst case into three segments and this happened in letters that belong to the classes: 15, 16, 17, 18, and in some types of the handwritten letters of classes 7, 10, 19, 20, 21, and 22. In addition, characters of classes 2, 3, 6, 11, 12, 13, 23, 24, 29, 32, 43, and 44, shown in Table 1,

are segmented into two segments in the most types of handwritings. Fig. 2 clarifies these cases of triple (three segments) and double (two segments) segmentation.

The main objective of the cooperation between the recognition model and the combination rules is to handle the over segmented letters introduced in Fig. 2.

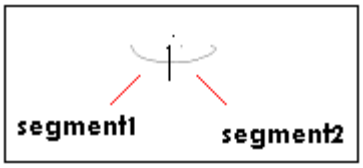

Fig. 3. Applied combination rules to double-segmented letter.

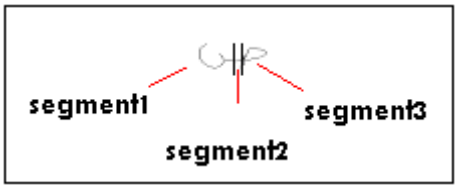

Fig. 4. Applied combination rules to triple-segmented letter.

First, the number of resulting segments was determined. Combination rules were applicable only when the number of segments is equal to or greater than two. Initially, the first two segments or three segments, if any, are passed to the neural network and recognized separately. If the first segment was recognized as a letter which can be a part of any other letter that may over-segmented into three parts, this first segment is combined with the second segment and the third segment and passed again to the neural network. If this combination was well recognized, then the final character will be the combination form of the three segments. Otherwise, only the first two segments are combined and passed to the neural network. If the first segment can be a part of any other letter that may over-segmented into two parts, then The final character will be the one with the higher recognition result of the first segment alone and its combination with the second segment. Finally, this process of combination and recognition repeated starting from the next unrecognized segment until all resulting segments are recognized as letters or combined into recognized letters. Fig. 3 illustrates an over-segmented letter 'ت', where the second segment was classified as class 5 and the combination form of the two segments was classified as class 6.

Fig. 4 also shows a triple-segmented letter 'ص', where the first segment was classified as class 43 and the combined form of the first, second, and the third segments was returned as class 22.

5 Experimental Results

More than 600 of over-segmented words were tested using the proposed recognition system aided by the above combination rules. The obtained results are very encouraging, as illustrated in Table 2. The percentage of correct segmentation increased from 86% before recognition and combination rules application to 95% after recognition and combination. On the other hand, the over-segmentation percentage decreased from 14% to 5% whereas the under-segmentation percentage remained constant.

Table 2. The Segmentation Results Using the Selected Sample.

	Before Recognition & Combination	After Recognition& Combination
% of correct segmentation	86%	95%
% of over segmentation	≈14%	≈5%
% of missing/wrong segmentation	0.3%	0.3%

The majority of the over-segmented characters are combined and recognized correctly. Fig. 5 depicts segmented words before and after recognition and combination rules application.

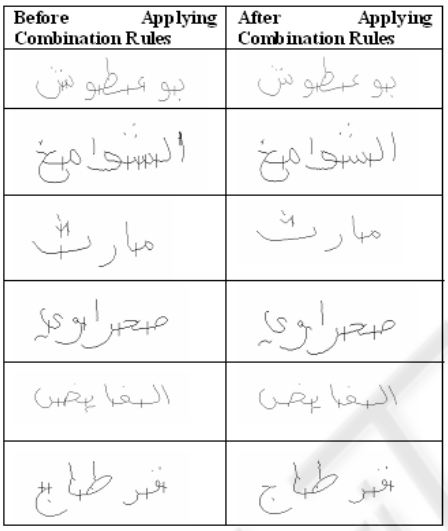

Fig. 5. Over-Segmented Words Before and After Applying Combination Rules.

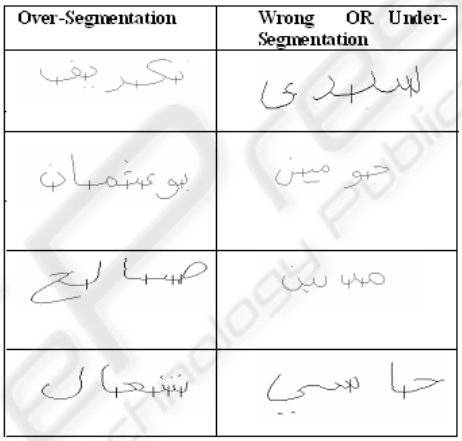

Fig. 6. The Problems of Character Misrecognition.

As demonstrated above, both double and triple segmented letters are combined correctly. The letter 'ش' which appears in the words 'الشوامخ' and 'بوعطوش' was segmented into three parts and each part is a candidate letter. However, the combination and recognition processes yield to one strong candidate letter instead of the three parts. A different example of double segmented letters can be observed in letters 'ث' and 'ي' that belong to words 'مارث' and 'صحراوي', respectively. Similarly, the combined segments had higher recognition rate, compared to each segment separately. The same case can be observed in the rest of examples.

Because the combination rules are strongly based on the recognition result of each segment, those which are misrecognized may not help in handling the over-segmented letters. This case appears in the letters of classes: 3, 6, 29, 32, 43, and 44, where the left segment of the over-segmented letter is wrongly recognized as letter alif 'ا', which belongs to class 1. One example is shown in Fig. 6, in the letter 'ف' of the word 'تكريف' and in the letter 'ن' in the word 'بوعثمان'.

The under-segmentation problem may also occur because of letter misrecognition. Adjacent letters that form shapes similar to those of classes 15, 16, 17, 18, 19, 20, 21 and 22 may wrongly be combined although they are correctly classified before combination. This scenario is very clear in Fig. 6, in the word 'سيدي', where the adjacent letters 'ي' and part of the letter 'س' were recognized and classified as letter 'س', and the remaining part of 'س' was classified as letter 'ل'. Similar case was also detected in the words 'مدنين' and 'حاسي', with different classification of the combined segments. In the word 'جومين', the two letters 'ي' and 'م' were combined and classified as letter 'ص'.

However, these results still fair because the combined segments form a body shape similar to that of an existing alphabetical letter. Moreover, the recognizer is not an interpreter to search for the meaning of the word based on the recognition results of its letters. Therefore, the obtaining outcomes are acceptable since no recent work could solve these situations.

6 Conclusions

In this paper, an effective segmentation method for Arabic handwriting was developed. The method used a multi-agent approach to segment words and relied on recognition to verify the validity of the candidate segmentation points. The use of an artificial neural network along with combination rules lead to a good treatment of the over-segmentation problem in Arabic handwritings. Furthermore, it achieved better results, when compared to similar works, by reducing the effect of under segmentation. This is attributed to the decision agent, which makes the proper decisions to identify the candidate segmentation points. The resulting segments are passed to the recognizer, which will invoke and apply the combination-rules agent on the unrecognized segments before passing it to the recognizer again. The experimental results (~ 95%) were very satisfactory and promising.

Our future direction will focus on improving this approach and including other styles of Arabic handwritings. On the improvement front, currently we are studying the use of SVM and HMM recent and relevant techniques.

References

1. M. Cheriet, N. Kharma and C-Lin Lui, C. Y.Suen, Character Recognition Systems: A Guide for Students and Practioners, John Wiley & Sons, Inc., 2007.
2. R. Bentrecia and A. Elnagar, Handwriting Segmentation of Arabic Text, International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA 2008), Innsbruck, Austria, Pages: 122-127, Feb. 13-15, 2008.
3. B. Al-Badr and S.A. Mahmoud, Survey and Bibliography of Arabic Optical Text Recognition, *Signal Processing*, vol. 41, pp. 49-77, 1995.
4. A. Amin, Offline Arabic Character Recognition: The State of the Art, *Pattern Recognition*, vol. 31, pp. 517-530, 1998.
5. A.S. Eldin and A.S. Nouh, Arabic Character Recognition: A Survey, *Proc. SPIE Conf. Optical Pattern Recognition*, pp. 331-340, 1998.

6. M.S. Khorsheed, Off-Line Arabic Character Recognition: A Review*, Pattern Analysis and Applications*, vol. 5, pp. 31-45, 2002.
7. B. Parhami and M. Taraghi, Automatic Recognition of Printed Farsi Texts, *Pattern Recognition*, vol. 14, pp. 395-403, 1981.
8. A. Amin and G. Masini, Machine Recognition of Multi-Font Printed Arabic Texts, *Proc. Int'l Conf. Pattern Recognition*, pp. 392-395, 1986.
9. A. Gillies, E. Erlandson, J. Trenkle, and S. Schlosser, Arabic Text Recognition System, *In Proceedings of the Symposium on Document Image Understanding Technology*, Annapolis, Maryland, 1999.
10. L. Hamami and D. Berkani, Recognition System for Printed Multi-font and Multi-size Arabic Characters, *The Arabian J. Science and Eng.*, vol. 27, pp. 57-72, 2002.
11. M. Dehghan, K. Faez, M. Ahmadi, and M. Shridhar, Handwritten Farsi (Arabic) Word Recognition: A Holistic Approach Using Discrete HMM, *Pattern Recognition*, 34(5):1057–1065, May 2001.
12. S.A. Al-Qahtani and M.S. Khorsheed, An Omni-Font HTK-Based Arabic Recognition System, *Proc. Eighth IASTED Int'l Conf. Artificial Intelligence and Soft Computing*, 2004.
13. S.A. Al-Qahtani and M.S. Khorsheed, A HTK-Based System to Recognize Arabic Script, *Proc. Fourth IASTED Int'l Conf. Visualization, Imaging, and Image Processing*, 2004.
14. B. Al-Badr and R. Haralick, A Segmentation-Free Approach to Text Recognition with Application to Arabic Text, *Int'l J. Document Analysis and Recognition*, vol. 1, pp. 147-166, 1998.
15. B. Al-Badr and R. Haralick, Segmentation-Free Word Recognition with Application to Arabic, *Proc. Int'l Conf. Document Analysis and Recognition*, pp. 355-359, 1995.
16. M.S. Khorsheed and W.F. Clocksin, Structural Features of Cursive Arabic Script, *Proc. British Machine Vision Conf.*, pp. 422-431, 1999.
17. A. Amin, Recognition of Printed Arabic Text Based on Global Features and Decision Tree Learning Techniques, *Pattern Recognition*, 33(8):1309–1323, August, 2000.
18. M. Pechwitz and V. Maergner, HMM-Based Approach for Handwritten Arabic Word Recognition Using the IFN/ENIT-Database, *In ICDAR IEEE Computer Society*, pp. 890–894, 2003.
19. R. C. Gonzalez and P. Wintz, Digital Image Processing, *2nd edition. Boston, Massachusetts: Addison-Wesley*, 1987.
20. Jan Teuber, Digital Image Processing, *Prentice Hall International Series in Acoustics, Speech and Signal Processing*, 1991.
21. P. Adibi, Farsi Handwritten Word Recognition Using a Continuous-Density Variable-Duration Hidden Markov Model, *Master of Science Thesis*, Computer Engineering Department, Amir Kabir University of Technology, Tahran, Iran, 2001.
22. R. Safabakhsh and P. Adibi, Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM, *The Arabian Journal for Science and Engineering,* Volume 30, Number 1B, 2004.