

Recognition-based Segmentation in Persian Character Recognition

Mohsen Zand, Ahmadreza Naghsh Nilchi, and S. Amirhassan Monadjemi

Abstract—Optical character recognition of cursive scripts presents a number of challenging problems in both segmentation and recognition processes in different languages, including Persian. In order to overcome these problems, we use a newly developed Persian word segmentation method and a recognition-based segmentation technique to overcome its segmentation problems. This method is robust as well as flexible. It also increases the system's tolerances to font variations. The implementation results of this method on a comprehensive database show a high degree of accuracy which meets the requirements for commercial use. Extended with a suitable pre and post-processing, the method offers a simple and fast framework to develop a full OCR system.

Keywords—OCR, Persian, Recognition, Segmentation.

I. INTRODUCTION

CHARACTER Recognition or Optical Character Recognition (OCR), is the process of converting scanned images of machine printed or handwritten text (numerals, letters, and symbols), into a computer format text (such as ASCII). OCR has been extensively used as the basic application of different learning methods in machine learning literature [1], [2]. This is the technology long used by libraries and government agencies around the world to make lengthy documents quickly available electronically.

A character recognition system can be either "online" or "offline." On-line is performed concurrently with the writing process. This is usually done through pen-based interfaces where the writer writes with a special pen on an electronic tablet. The online system does not require thinning and/or skeletonization process, as the pen itself is one pixel thick.

Off-line recognition, on the other hand, is performed after the writing has been performed. The basic unit is a bitmap, and, therefore, thinning and skeletonization processes are extensively required during the preprocessing stage.

In both online and offline OCR systems, there are three main approaches for automatic understanding of cursive

scripts including Persian scripts, namely holistic, segmentation-based, and recognition-based methods [3].

In the first approach, each word is treated as a whole and the recognition system does not consider it as a combination of separable characters. Very similar to the speech recognition systems, in almost all significant results obtained from holistic methods, Hidden Markov Models have been used as the recognition engine [4], [5]. The second approach, which has done extensively by researchers in OCR systems of such scripts, segments each word which contains characters as the building blocks, and then it try to recognize each character as the next step in the process.

In comparison, the first approach usually outperforms the second, but it still needs a more detailed model of the language, which its complexity grows as the vocabulary set gets larger. In addition, in this method, the number of recognition classes is far more than classes in segmentation-based methods.

Recently, there has been a new third approach which tends to hybrid both methods, called segmentation-by-recognition methods. The approach is related to some early ideas of the OCR of the isolated Roman characters. This approach incorporates the segmentation and recognition systems to obtain overall automatic understanding process of the scripts.

One of the most difficult problems in Persian optical character recognition process is the handle of the Persian texts' cursiveness. Thus while the segmentation approach is relatively simple in printed Roman texts, it is still an open question in Persian. However, the most reported segmentation methods OCR in Persian OCR systems to date include this approach. We believe this is the main source of recognition errors.

In this paper, the third approach namely, the segmentation-by-recognition method is considered for Persian OCR systems in order to reduce this error.

II. THE CHARACTERISTIC OF PERSIAN SCRIPT

In this section, we will briefly describe some of the main characteristics of Persian script to point out the main difficulties which an OCR system should overcome.

As one of the main properties, the script consists of separated words which are aligned by a horizontal virtual line called "baseline". Words are separated by long spaces and each word consists of one or more isolated segments each of

Mohsen Zand is a faculty member at the department of computer engineering, Islamic Azad University of Doroud, Doroud, Iran (e-mail: zand.mohsen@gmail.com).

Ahmad R. Naghsh Nilchi is an assistant professor at the department of computer engineering, faculty of engineering, University of Isfahan, Isfahan, Iran (e-mail: nilchi@eng.ui.ac.ir).

S. Amirhassan Monadjemi is an assistant professor at the department of computer engineering, faculty of engineering, University of Isfahan, Isfahan, Iran (e-mail: monadjemi@eng.ui.ac.ir).

them are called sub-word. On the contrary sub-words are separated by short spaces and each sub-word includes one or more characters. If one sub-word has more than one character, each of them will be connected to its neighbors along the baseline.

As each Persian character has two to four different forms, this extends the classes to be recognized from 32 to 120. Fig. 1 shows the character set of Persian script which clearly illustrates that the appearance of Persian character varies according to its position in a word or sub-word [6], [7].

Both typed and hand-written Persian scripts are cursive and are read from right to left. Due to the cursive nature of the script, we can either recognize a word at a time or segment a word into characters and then recognize the characters.

The first case seems to be impossible and not feasible due to the numerous numbers of words in a language. However, if the second case is used, research has been practically proved that the segmentation of a cursive word is a very difficult problem. However the segmentation is a crucial step in Persian OCR systems [8].

We have also noticed that some Persian words may be horizontally overlapped with others in a document. An example is given in Fig. 2. This feature causes the traditional segmentation method using projection profile not applicable in this situation and it brings out the word segmentation problem.

Some other characteristics of Persian script are summarized below.

None	Detached	Initial	Medial	Final	None	Detached	Initial	Medial	Final
Alef	ا	ا	ا	ا	Sad	ص	ص	ص	ص
Be	ب	ب	ب	ب	Zad	ض	ض	ض	ض
Pe	پ	پ	پ	پ	Ta	ط	ط	ط	ط
Te	ت	ت	ت	ت	Za	ظ	ظ	ظ	ظ
Se	ث	ث	ث	ث	Ein	ع	ع	ع	ع
Jim	ج	ج	ج	ج	Ghein	غ	غ	غ	غ
Che	چ	چ	چ	چ	Fe	ف	ف	ف	ف
He	ح	ح	ح	ح	Qaf	ق	ق	ق	ق
Khe	خ	خ	خ	خ	Kaf	ک	ک	ک	ک
Dal	د	د	د	د	Gaf	گ	گ	گ	گ
Zal	ذ	ذ	ذ	ذ	Lam	ل	ل	ل	ل
Re	ر	ر	ر	ر	Mim	م	م	م	م
Ze	ز	ز	ز	ز	Nun	ن	ن	ن	ن
Zhe	ژ	ژ	ژ	ژ	Vav	و	و	و	و
Sin	س	س	س	س	Hed	ه	ه	ه	ه
Shin	ش	ش	ش	ش	Ye	ی	ی	ی	ی

Fig. 1 The Persian Alphabet

Most characters (20 out of 32) have a dot, two dots, or zigzags associated with the character and they are located either above, below, or inside the character.

Most characters share similar shape with others. The position or number of dots in the character makes the only difference.



Fig. 2 An example of overlapped Persian words

Some characters can only appear at the beginning or at the end of a word or sub-word.

A Persian word could have one or more sub-words. This is due to the fact that some characters are not connectable from the left side with the succeeding character.

III. THE OUTLINE OF THE APPROACH

The implemented Persian OCR system involves five image processing techniques which are the image acquisition, the preprocessing, the segmentation, the feature extraction and the classification. However, as a recognition-based character segmentation technique is used, a feedback loop is linked between the output of the classification stage and the input of the character fragments combination stage.

The block diagram of the recognition-based Persian OCR system is shown in Fig. 3.

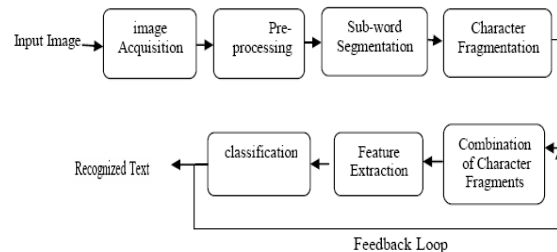


Fig. 3 The recognition-based Persian OCR system

Image acquisition and preprocessing are the two relatively simple stages. Image acquisition is at the image representation level of pattern recognition (PR). It is the process of acquiring a digitized representation of a document or an article to be recognized.

Preprocessing of the image is done to prepare it for other stage. It increases the accuracy of the recognizing algorithm by enhancing some of the features and eliminating some of the inconsistencies. Preprocessing is at the image-to-image transformation level. It is the process of compensating a poor-quality original and/or poor-quality scanning [9], [10].

The image is then ready for segmentation. The projection profile method is employed to extract lines from the document. As mentioned earlier in Section 2, Persian words

may horizontally overlap with others, therefore a word segmentation method is developed to solve this problem. The algorithm is described in [11].

Since the word segmentation method can accurately separate horizontally overlapping Persian words/sub-words efficiently, it is a real-time process. It is hard to develop dissection rules for a cursive script. Therefore, we fragmented Persian words using their structural properties and connectivity points. We then recognized characters by combining fragments.

This technique bypasses the segmentation step so that we do not have to worry about determining the actual character segmentation points.

IV. SEGMENTATION

Segmentation is a crucial step of OCR systems as it extracts meaningful regions for analysis. A poor segmentation process produces mis-recognition or rejection. It is especially important for Persian OCR systems due to the cursive nature of Persian script and the fact that some Persian words overlap vertically. Page layout analysis and character separation are used to segment sub-words from the preprocessed image.

In the page layout analysis process, a horizontal projection profile of the document image is plotted. The segmentation between lines of text is determined by scanning through the profile from the first row. If the difference of the number of black pixels between two rows is larger than a predefined threshold, a new line of text is indicated (depicted in Fig. 4). The next large variation in the number of black pixels between another two rows indicates the bottom of the line.

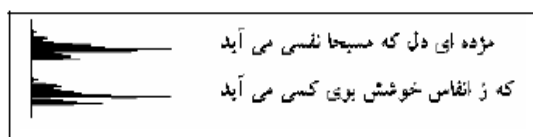


Fig. 4 Using horizontal histogram to separate lines

Sub-words are segmented from a line segmented image in a similar method, except that a vertical projection profile is plotted instead [12]. An example is given in Fig. 5.

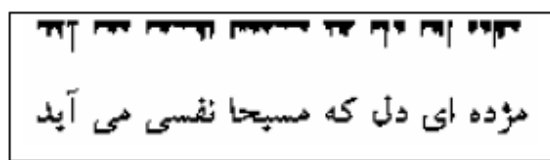


Fig. 5 Using vertical histogram to separate sub-words

V. CHARACTER FRAGMENTATION

The purpose of this step is to produce a sequence of tentative character segmentation lines. Although these lines do not necessarily segment characters from a word correctly, the location and the number of them would directly affect the accuracy and speed of the OCR system [11].

It is important to note that this step only produces a

sequence of fragments, while the segmentation of characters is confirmed at the classification stage.

In this Character fragmentation method, we search for the preliminary segmentation paths.

The algorithm for searching the preliminary segmentation path of Persian words:

- 1) Average fragment size for the current address is calculated, by scanning for segregated characters and noting their width and height.
- 2) If a column exists, check its pixel density. Else go to Step 10.
- 3) If the pixel density is zero, then segmentation point found. Bypass ensuing columns with pixel density zero, until the beginning of the next fragment encountered. Go to Step 2.
- 4) If previous or next column's pixel density is greater than current column, go to Step 5. Else return to Step 2.
- 5) Calculate how many columns have been passed since last segmentation point.
- 6) If the number of columns is smaller than the average size of the fragment, go to Step 2.
- 7) If any ensuing columns have a pixel density of zero, go to Step 2.
- 8) Check if average pixel density of previous and ensuing columns is greater than that of the proposed point.
- 9) If Step 8 is true, then segmentation point found. Repeat by going to Step 2.
- 10) End of algorithm.

An example is given in Fig. 6.



Fig. 6 Character segmentation lines

VI. FEATURE EXTRACTION

After an image has been segmented into regions, it is ready to enter the next level that is, the feature extraction stage.

The end result of the image acquisition, preprocessing, segmentation and character fragmentation is a matrix of numbers that represents a character fragment in some way. In the general case, however, the matching of these numbers to a template may be too time consuming and not flexible enough. Therefore, feature extraction is needed. Structural features of each character fragment are extracted in this method.

VII. CLASSIFICATION

The classification process is carried out at the final stage to recognize the character. It assigns an input character to one of many pre-specified classes which are based on the extracted features and their analysis.

For the classification process, ANN (Artificial Neural Network) was used in the proposed method. It had utilized to

search for the identity of a character [13].

To recognize characters, sub-words are first fragmented into a sequence of character fragments by the method described in Section 5.

In this method, each fragment is numbered from right to left. During the recognition process, the first fragment is fed into the feature extraction stage to determine the concentrated codes.

These codes are then input to the ANN to find the best match. In order to minimize the confusion of character fragments with characters and to save search time, there are four databases. According to the position of a fragment(s) in a word, the corresponding database is used to search for the best match. For example, if a tentative character is formed by the combination of the first and second fragment, the database for the beginning form characters is used. If this tentative character could not be recognized, a signal is fed back to the character combination process to combine the first three character fragments (refer to Fig. 7).

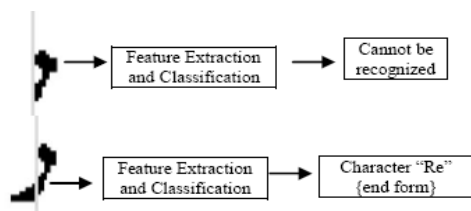


Fig. 7 Right-to-left feedback loops for recognition of the Persian character

The above processes are repeated until a character is recognized. If a character is recognized after the combination of the first n fragments, then the feedback loop will start again at the $(n+1)$ th fragment. The above feedback loop occurs twice for each word. The first is with the fragment combination directed from right to left of the word. If not all characters in that word are recognized, the second feedback loop proceeds. This time, the fragment combination is directed from left to right of the word (refer to Fig. 8). The results from these two feedback loops are combined to form the final recognition results.

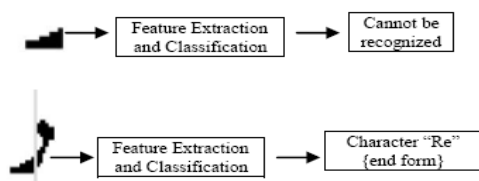


Fig. 8 Left-to-right feedback loops for recognition of the Persian character

Refer to the Persian character set shown in Fig. 1, we can see that some characters look similar to a part of some other characters, e.g., the middle form of Be, Pe, Te, and The, look

similar to a fragment of the middle form of Sin, Nun and Shin. As there is no exact character segmentation point provided, confusion between these characters may occur and lead to mis-recognition.

Since we have to understand that recognized or non-recognized is the guideline for combining character fragments, any mis-recognition in the middle of the words will affect the lot. That is the reason why the second directed feedback loop was used. It is used to compensate for some errors occurring due to mis-recognition in the first trial.

VIII. CONCLUSION

Segmentation introduces the most serious problem in the development of cursive script OCR system including Persian language scripts. In order to overcome this problem, we use a newly developed word segmentation method and a recognition-based segmentation technique. That is, we fragmented Persian words using their structural properties and connectivity points. We then recognize characters by combining fragments. This technique bypasses the segmentation step so that we do not have to be worried about determining the actual character segmentation points. The method is simple to implement and does not require lengthy numerical computations.

The drawback of this method is in the classification stage. Even though we have performed right-to-left and left-to-right feedback recognition, whenever there is a character in a word that could not be recognized, the rest of the characters in the word are not recognized properly as well. It seems, by offering a better classification method, this problem could be solved.

Comparing the recognition accuracy of our method with others shows that the recognition accuracy has increased, significantly. In addition, this recognition-based model seems to be more suitable to other cursive scripts including handwritten Latin.

REFERENCES

- [1] A. Amin, "Off line Arabic character recognition – a survey", Proceedings of the International Conference on Document Analysis and Recognition, vol. 2, pp. 596-599, 1997.
- [2] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient based learning applied to document recognition", Proceedings of the IEEE, vol. 86, no. 11, IEEE, pp. 2278- 2324, USA , 1998.
- [3] B. Al-Badr, R. Haralick, "Segmentation-free word recognition with application to Arabic", Proceedings of the Third International Conference on Document Analysis and Recognition, Part vol. 1, IEEE Comput. Soc. Press., pp. 355-359, Los Alamitos, CA, USA, 1995.
- [4] I. Bazzi, R. Schwartz, J. Makhoul, "An omnifont open vocabulary OCR system for English and Arabic", IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 21, no. 6, IEEE Comput. Soc., pp. 495-504, USA, 1999.
- [5] A. Hassin, Tang, Xiang-Long, Liu, Jia-Feng, Zhao-Wei, "Printed Arabic character recognition using HMM", Journal of Computer Science & Technology, vol. 19, no. 4, Science Press, pp. 538-543, China, 2004.
- [6] I. S. Abuhaiba, S. A. Mahmoud, and R. Green, "Cluster Number Estimation and Skeleton Refining Algorithm for Arabic Characters", The Arabian Journal for Science and Engineering, vol. 16, no. 4B, pp. 519-530, 1991.

- [7] K. Jambi, "Arabic Character Recognition", Many Approaches and One Decade., *Die Arabic Journal for Science and Engineering*, vol. 16, no. 4B, pp. 501-509, 1991.
- [8] A. Cheung, M. Bennamoun, and N. W. Bergmann, "Implementation of A Statistical Based Arabic Character Recognition System", *TENCON97*, pp. 531-534, Brisbane, Australia, 1997.
- [9] K. R. Castleman, "Digital Image Processing", Prentice-Hall Signal Processing Series, Prentice-Hall Inc., USA, 1979.
- [10] R. C. Gonzalez, P. Wintz, "Digital Image Processing", 2nd Edition, Addison-Wesley Publishing Company, California, 1987.
- [11] A. Cheung, M. Bennamoun, and N. W. Bergmann, "A New World Segmentation Algorithm for Arabic Script", *DICI'A'97*, pp. 431-435, Auckland, New Zealand, 1997.
- [12] B. Timsari, "Character recognition in typed Persian words", a morphological approach, M.S. thesis, Isfahan Univ. of Tech., Iran(1992)
- [13] R. J. Schalkol, "Pattern Recognition: Statistical, Structural and Neural Network", Wiley, New York, 1992.

Mohsen Zand was born 1982, in Tehran, Iran. He got his BSc in software engineering from University of Isfahan, Isfahan, Iran, in 2005. He is now an MSc student at the Azad University of Najafabad, Isfahan, Iran. He also is a faculty member at the Computer Engineering Department, Azad University of Doroud, Doroud, Iran.

Ahmad R Naghsh Nilchi, PhD, received his B.S. and M.S., and PhD degrees from Electrical and Computer Engineering Department in 1988, 1989, and 1996, respectively, all from the University of Utah, Salt Lake City, Utah, USA. He has been awarded several research grants from distinguished research institutions including U.S. National Science Foundation and has completed several research projects for Iranian industries. He also is the chief editor of the Iranian Journal of Engineering Sciences. His research interests include Image and signal processing, data hiding, character recognition, computer graphics, and complex numerical analysis and applications. Since 1997, he has been a faculty member at the Computer Engineering Department, University of Isfahan, Isfahan, Iran.

Seyed Amirhassan Monadjemi, PhD, was born in 1968, in Isfahan, Iran. He got his PhD in computer engineering, pattern recognition and image processing, from University of Bristol, Bristol, England, in 2004. He is now working as an assistant professor at the Department of Computer, Faculty of Engineering, University of Isfahan, Isfahan, Iran. His research interests include pattern recognition, image processing, and human/machine analogy.