International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:4, 2007

A Persian OCR System using Morphological Operators

M. Salmani Jelodar, M.J. Fadaeieslam, N. Mozayani, M. Fazeli

***Abstract*—** Optical Character Recognition (OCR) is a very old and of great interest in pattern recognition field. In this paper we introduce a very powerful approach to recognize Persian text. We have used morphological operators, especially **Hit/Miss** operator to descript each sub-word and by using a template matching approach we have tried to classify generated description. We used just one font in two different sizes to verify our approach. We achieved a very good rate, up to 99.9%.

***Keywords*—**APersian Optical Character Recognition.

I. INTRODUCTION

THE area of machine-printed and hand-written character and word recognition has been the subject of many research studies in the past few years [1][2][3][4]. Several algorithms have been proposed for the recognition of widely used characters, in both printed and cursive forms. However, still there are unsolved problems in this area. Persian characters used as the script for languages such as Farsi, Arabic, and Urdu, are among those characters developing a recognition system for them have not received so much attention [1][8][12]. Since the development of any object recognition scheme is a direct consequence of the characteristics of the object being recognized, it is not possible to apply directly many of the recognition algorithms proposed for other classes of characters to Persian characters.

Our objective in this paper is to present and demonstrate an approach, not just a technique, for Persian character recognition, based on the morphological image processing methods. We start with a brief review of some important Persian text's features and then introduce in brief morphological transformations. Then we explain in some details our proposed method and follow that with experimental results.

II. PERSIAN TEXT FEATURES

Persian character recognition is different from Latin character recognition. It is a cursive script and it has many different features.

Manuscript received January 21, 2005.
Mehdi Salmani Jelodar, Mohammad Javad Fadaei Eslam and Mahlagha Fazeli are with the Sepanta Robotic Research Foundation. (e-mails: m.salmani@srrf.net, Mj.Fadaei@srrf.net, M.Fazeli@srrf.net)
Naser Mozayani is with Iran University of Science and Technology. (email: Mozayani@iust.ac.ir)

We can classify Persian language features in several items:

1- Persian text is written from right to left.
2- Farsi has 32 characters out of which 18 have 1 to 3 points which maybe locate below (like "پ"), above (like "ث") or in the middle of the character (like "چ"). Some of them maybe have some other vowels (like "ء" and " ً ").
3- Persian is a cursive script. Characters are connected and make a component. These components are called "sub-words". A single isolated letter is considered as the extreme case of a sub-word. A word may have several sub-words for example "مهدی" is a word and has two sub-words "مهد" and "ی".
4- In Persian, depending on its position in a sub-word, a letter may take different shapes. Although there are only 32 letters in Persian alphabet, the total number of different classes to be recognized sums up to 127 (Table 1.). Most of the letters have dots above, below, or inside them, number of which is variable between one to three. There are letters whose only difference is the number and/or location of their dots. Ignoring the dots reduces the number of classes to 66. If we consider dots, signs (like "ء" and …) and numbers we have 157 different classes.
5- In Persian/Arabic text there is a base line which usually has more black pixels. (Fig 1.).

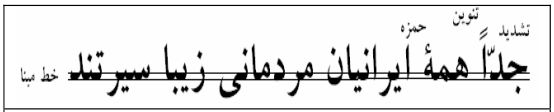

Fig. 1. A Persian text and its base line

There are other features which are not very important.

Some of these features make Persian character recognition very hard and complex. Because of the connectivity between Persian characters, its recognition is very hard and most of errors occur in the segmentation phase.

III. PERSIAN CHARACTER RECOGNITION ALGORITHM

The recognition of characters in Persian printed documents is performed in four stages as depicted in Fig. 2 (without considering output stage).

Table 1- Persian Charcters

End	Middle	First	Isolated
ـا			ا آ
ـب	ـبـ	بـ	ب
ـپ	ـپـ	پـ	پ
ـت	ـتـ	تـ	ت
ـث	ـثـ	ثـ	ث
ـج	ـجـ	جـ	ج
ـچ	ـچـ	چـ	چ
ـح	ـحـ	حـ	ح
ـخ	ـخـ	خـ	خ
ـد			د
ـذ			ذ
ـر			ر
ـز			ز
ـژ			ژ
ـس	ـسـ	سـ	س
ـش	ـشـ	شـ	ش
ـص	ـصـ	صـ	ص
ـض	ـضـ	ضـ	ض
ـط	ـطـ	طـ	ط
ـظ	ـظـ	ظـ	ظ
ـع	ـعـ	عـ	ع
ـغ	ـغـ	غـ	غ
ـف	ـفـ	فـ	ف
ـق	ـقـ	قـ	ق
ـک	ـکـ	کـ	ک
ـگ	ـگـ	گـ	گ
ـل	ـلـ	لـ	ل
ـم		مـ	م
ـن	ـنـ	نـ	ن
ـو			و
ـه	ـهـ	هـ	ه
ـی	ـیـ	یـ	ی

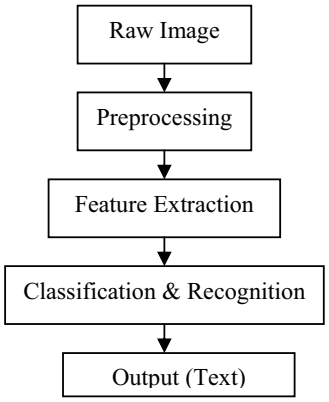

Figure 2-A Persian Character Recognition System

In the input stage reading Raw Images is done using Bitmap format.

In *preprocessing* stage we have divided lines and after that we set apart sub-words. After that we count and eliminate dots. At the end of this stage thinning sub-words is done. The steps are described as follow:

A. Line Segmentation

Between each two lines there is a free space and it is used to segment lines. We use horizontal histogram of documents to separate lines (depicted in Fig. 3) [10].

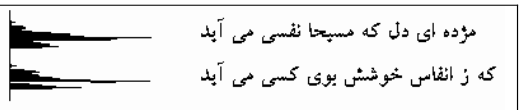

Figure 3- Using horizontal histogram to separate lines

B. Sub-word Segmentation

There is a free space between each two sub-words as well and we used this feature to segment sub-words. We use vertical histogram of lines to separate sub-words (depicted in Fig. 4).

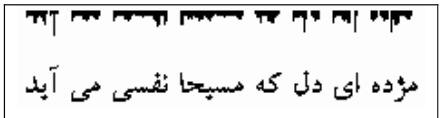

Figure 4- Using vertical histogram to separate sub-words

C. Removing Dots

Dot counting and removing will help us to reduce our classes and it makes recognition stage simpler [5] [6]. We count number of black pixels, if it is less than or more than a constant value, and located in an expected location, it is considered to be a dot and it is removed.

D. Thinning

Thinning is one of the most important steps in preprocessing stage [11]. We have used the sequential thinning method based on morphological hit/miss transformation [10]. This algorithm has two cycles. The first cycle is an iterative parallel process in which, each iteration is consists of four passes. Each pass uses a structuring element to remove border points from a given direction. In other words, points that satisfy the hit/miss transform of the structuring element are removed. The structuring elements for the first cycle used to remove border points from the four minor directions are shown in
Figure 5 as first cycle. The second cycle is another iterative process designed to remove border points from the major directions. The four structuring elements used for this purpose are shown in Figure5 as second cycle.

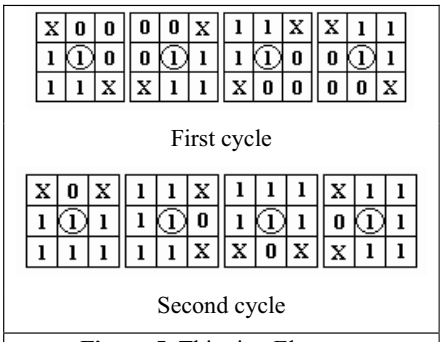

Figure 5. Thinning Elements

Let S be the image of the sub-word, and I be its skeleton. If we index the iteration process of the first and second cycles by m and n respectively, the two cycles of the thinning process can be expressed by

$$T = \{S \Diamond D\}_m$$

$$I = \{T \Diamond E\}_n$$

Where $\{S \Diamond D\}_m$ means the thinning of S by the four structuring elements $\{Di\}_{i=1}^{4}$ for m iterations, ($\{T \Diamond E\}_n$ is defined similarly), and the thinning operation is defined based on hit/miss transformation as $U \Diamond V = U - (U \otimes (V, V^C))$. After a finite number of iterations, depending on the maximum thickness of *S*, *S* converges to the skeleton *I*.
Trying to save in the computation time, the thinning process is interrupted before the image converges to its skeleton.
Depending on the thickness of characters, after a few number of alternative iterations of both cycles of the thinning process, the image converts to a form appropriate for extracting features used for recognition (Figure 6).

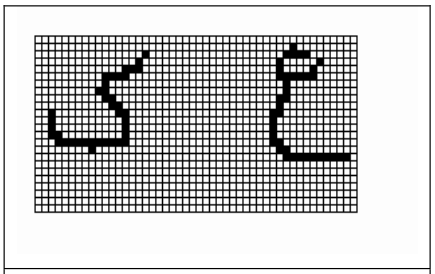

Figure 6. Thinned Persian characters.

Feature extraction is a very important stage in each optical character recognition system [7][8][9]. Simple geometric patterns such as horizontal and vertical straight lines, right angled corners, and T junctions are the features we look for in the thinned image of the sub-words. After thinning a printed Persian character, regardless of its form, it can always be decomposed into a set of primitive patterns. A combination of some appropriately chosen subset of these patterns uniquely defines that character. Feature extraction is done by an exhaustive search process using the hit/miss operator with a complete set of structuring elements corresponding to the different geometric patterns of our interest (Fig. 7).

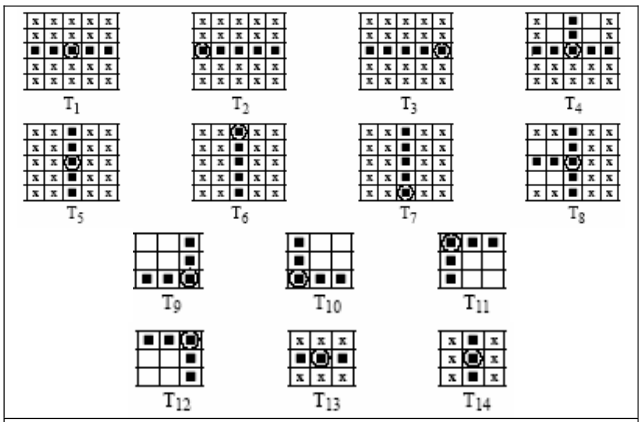

Figure 7. Structuring elements used for feature extraction

Final stage is *Decision and Classification.* Features extracted in the previous stage, along with the knowledge about the number and position of dots provide the required information for identification of the letters in a sub-word. A unique description has been assigned to each letter based on its constituting primitive patterns and their relative positions, plus the number and location of its dots. Each description forms the recognition rule for a letter. If the configuration of patterns found in a character matches one of the descriptions, the corresponding rule identifies the character. Whenever the algorithm is trained for a new font, the rules are modified to cover possible new descriptions of letters.

IV. Experimental Results

The algorithm was tested on Lotus font, which seems to be the most common Persian font. We used this font in two different sizes to recognize the characters. The average recognition rate for the implementation of the method running on a common PC (Pentium III 800 MHz) was 300 characters per second. Using a test set containing 3000 words (15000 characters) an accuracy rate of 99.9% was measured for this algorithm.

V. Conclusion and Future Works

A method for recognition of Persian characters in machine printed documents is developed based on the morphological hit/miss transformation. The capabilities of this operator in detecting patterns with specific geometric properties in the image, is used appropriately to accomplish different essential tasks in a pattern recognition process. In this method segmentation of word characters is not required for their classification. The characters are recognized at their positions in the word. This algorithm is trainable for different fonts and it can be extended to include digits too.

Morphological operators are very powerful but they are very sensitive to noise. And to reduce this sensitivity we can use skeleton of image instead of thinned image. In our future work we will use fuzzy morphological operator to make our approach more powerful and reliable.

References

[1] Badr Al-Badr and Saberi A.Mahmoud, "Survey and bibliograghy of Arabic optical text recognition",Elsevier Signal Processing,1995, pp.49-77.

[2] seera, J.,Image Analysis and Mathematical Morphology", Acrdemic Press, NewYork, 1982

[3] J.W. Smith and Z. Merali, "Optical character recognition", The British Library, Wetherby, West Yorkshire LS23 7BQ, UK, 1985.

[4] E.M. Welch, "Can you read this? OCR software", MacUser, Vol. 9, No. 8, November 1993, pp. 169-178.

[5] R.Azmi and A.Kabir "A new segmentation technique for omnifont Farsi text",Elsevier Pattern Recognition Letters, 2001, pp. 97-104.

[6] B. Parhami and M. Tarighi, "Automatic recognition of printed Farsi text", Pattern Recognition, Vol. 14, No. 1,1981, pp. 1-6.

[7] H. Almuallim and S. Yamaguchi, "A method of recognition of Arabic cursive handwriting," IEEE Trans. Patt. Anal. Machine Intell., vol. PAMI-9, no. 5, Sept. 1987.

[8] T. El-Sheikh and R. Guindi, "Computer recognition of Arabic scripts," Patt. Recogn., vol. 21, no. 4, pp. 293-302, 1988.

[9] M. El-Wakil and A. Shoukry, "On-line recognition of handwritten isolated Arabic characters," Patt. Recogn., vol. 22, no. 2, pp. 97-105, 1989.

[10] B. Timsari, *Character recognition in typed Persian words: a morphological approach,* M.S. thesis, Isfahan Univ. of Tech., Iran, 1992.

[11] B. K. Jang and R. T. Chin, "Analysis of thinning algorithms using mathematical morphology,"IEEE Trans. Patt. Anal. Machine Intell., vol. PAMI-12, no. 6, pp. 541-551, June 1990.

[12] A. Amin and G. Masini, "Machine recognition of multifont printed Arabic texts," in *Proc. 8th Int. Conf. Patt. recogn.,* pp. 392-395, Paris, 1986.