A Front-end OCR for Omni-font Persian/Arabic Cursive Printed Documents

Ramin Mehran
1. Department of Electrical Engineering, K.N.Toosi Univ. of Tech., Tehran, Iran
2. Paya Soft co., Tehran, Iran
rmehran@ee.kntu.ac.ir

Hamed Pirsiavash
1. Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran
2. Paya Soft co., Tehran, Iran
h_pirsiavash@mehr.sharif.edu

Farbod Razzazi
Paya Soft co., Tehran, Iran
razzazi@payasoft.com

Abstract

Compared to non-cursive scripts, optical character recognition of cursive documents comprises extra challenges in layout analysis as well as recognition of the printed scripts.

This paper presents a front-end OCR for Persian/Arabic cursive documents, which utilizes an adaptive layout analysis system in addition to a combined MLP-SVM recognition process. The implementation results on a comprehensive database show a high degree of accuracy which meets the requirements of commercial use.

1. Introduction

Optical character recognition (OCR) has been extensively used as the basic application of different learning methods in machine learning literature [1, 2]. Consequently, there are also a large number of commercial products available in the market for recognizing printed documents. However, the majority of the efforts are focused on western languages with Roman alphabet and East Asian scripts. Although there has been a great attempt in producing omni-font OCR systems for Persian/Arabic language, the overall performance of such systems are far from perfect. Persian written language, which uses modified Arabic alphabet, is written cursively, and this intrinsic feature makes it difficult for automatic recognition.

An essential part of a document understanding process, which seems trivial in human's character recognition, is how we perceive the written text parts and how we distinguish between different sections of a document. To eye of an educated person, the layout of a document is discerned once it came to sight but to a machine, this seemingly simple task will become burdensome. To a large extend, the layout of written materials appear in rectangular columns with all text and shapes aligned to the margins. This arrangement is mostly referred to as Manhattan style documents. In addition to this style, there are more sophisticated styles where printed text and figures in a document are not aligned. Thus, majority of efforts for layout analysis of documents are focused on Manhattan style, but recently, there is a growing attention toward the complicated styles as well [3]. Persian/Arabic printed documents are treated in the same way in layout analysis however; there is an absence of solid resources in the literature for these languages.

Once the layout of a document is extracted, the system has access to each line of text. Thereafter, the problem of understanding of the script reduces into recognition of the words. There are two main approaches to automatic understanding of cursive scripts: holistic and segmentation-based [4]. 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 of holistic methods, Hidden Markov Models have been used as the recognition engine [5, 6]. The second strategy, which owns the majority in the literature, segments each word to containing characters as the building blocks, and recognizes each character then.

In comparison, the first strategy usually outperforms the second, but it 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 similar number in segmentation-based methods. Recently, there is also a trend toward hybrid methods which incorporates the segmentation and recognition systems to obtain overall results; these methods are usually called segmentation-by-recognition [7, 8].

One of the main concerns of designing every OCR system is to make it robust to the font variations. Thus, successful examples are omni-font recognition systems with ability to learn new fonts from some tutor. In holistic methods, as the character recognition problem is viewed from a different perspective, and the system collectively uses learning mechanisms for few connected characters, the transformation of the system into an omni-font learning system would be smooth. On the other hand, the segmentation-based systems mainly use learning methods only in recognition process, and to the best of our knowledge, the learning systems are never used for the segmentation process in the literature [9]. Usually, human recognizes unfamiliar words by segmenting them and recognizing each character separately to understand the whole word. With this perspective, in this research, the whole task is broken down into two separate learning systems to gain from reduction of complexity in hierarchy as well as adaptability of learning systems.

The layout of this paper is as follows: Section 2 emphasizes on the characteristics of Persian script that were crucial for the design of OCR systems. In Section 3, we will discuss the proposed algorithm, which includes the layout analysis, segmentation, and recognition modules in separate subsections. In Section 4, implementation details and results are discussed which is entailed with conclusive remarks and acknowledgements.

2. Some notes on Persian/Arabic script

In this section, we will briefly describe some of the main characteristics of Persian/Arabic 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 them is called Piece of a Word (PAW). On the contrary PAWs are separated by short spaces and each PAW includes one or more characters. If one PAW has more than one character, each of them will be connected to its neighbors along the baseline. Figure 1 shows a sample Persian/Arabic script, where *a* represents the space between two different words, and *b* is the short space between PAWs of a word.

In Figure 2, the first PAW on the right comprises three characters and the second one, on the left, consists of only a single character, and *p* denotes the pen width value which is heuristically equal to the most frequent value of the vertical projection in each line.

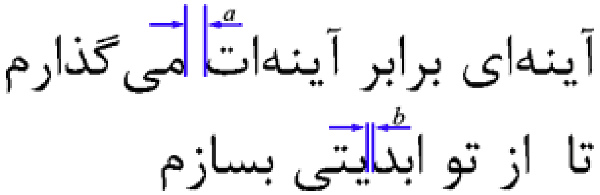


Figure 1. Sample of Persian Script and virtual baseline shown only for demonstration, where a represents the space between two different words, and b is the short space between PAWs of the first word.

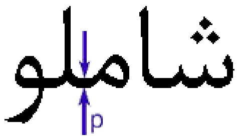


Figure 2. An example of a Persian word consists of two PAWs, where p represents pen width. The red line is the visualization of the imaginary baseline.

3. Proposed algorithm

The overall block diagram of the system is presented in Figure 3 which depicts layout analysis, post-process, and natural language processing (NLP) subsystems in addition to recognition and segmentation blocks. This paper presents the design of the layout analysis system in addition to the segmentation, feature extraction, and recognition sections (Figure 3). The design of the NLP module is out of scope of this paper.

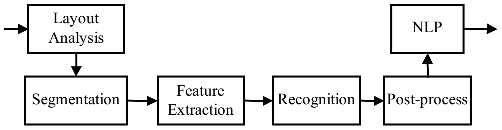


Figure 3. Overall block diagram

With respect to the popularity of Manhattan style in Persian documents, this research was focused on documents with aligned columns without any graphical figure inside; therefore, the automatic exclusion of the graphical material from text is left for future works. The designed method includes a two-step skew detection with use of a combination of basic Hough transform method [10] and the method proposed in [11] that uses fuzzy run-lengths. The layout analysis section will process a document and adaptively segments it into paragraphs and then into the separated lines.

In the proposed system, the segmentation-based approach is exploited, and some measures are considered to overcome the main weaknesses of it. The best results are achieved with aid of artificial neural networks (ANN) for performing segmentation with some extended features (Section 3.2). In recognition section, we obtained a definite set of features from each segmented symbol, which was fed to a support vector machine (SVM) classification engine to obtain the recognized symbol. Using large margin classifiers enables us to achieve high recognition rates which are in coherence with the best results in the literature [2].

We also decomposed each character of Persian script to more primitive symbols called graphemes. This novel decomposition has decreased the complexity of the recognition and segmentation procedures and has improved the overall result. Few different characters could share a single grapheme, and additionally, several joint graphemes could build a single character. In addition, Persian language includes many characters which the only difference they have is the number of dots and placement of them.

To finalize the character recognition task, a post-processing section is implemented to combine the result of grapheme recognition and the number of dots. Besides, this section corrects some common grapheme recognition errors using an embedded confusion matrix. Figure 4 shows the combination of grapheme recognition and post-processing processes with dot recognition module.

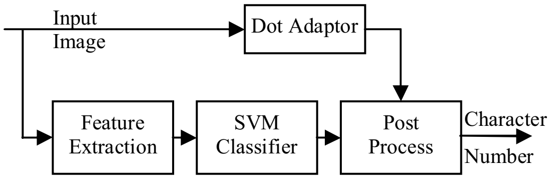


Figure 4. Grapheme recognition subsystem is combined with dot recognition modules and post-processing procedures to recognize characters.

Before proceeding further, we provide concepts of some frequently used terms in this paper for clarification:

Grapheme: In this research, we refer grapheme to any graphical image that would be a character or a part of it which acts as a fundamental building block of words. This resembles the concept of phonemes in speech, but we don't directly choose them in relation to real phonemes.

Pen tip: The vertical position of the pen in the skeleton of a PAW image.

Junction points: The horizontal position of the grapheme boundary. Thus, cutting the word at junction points results separated graphemes.

3.1. Layout analysis

The layout analysis section is responsible for segmenting document images into lines. The input of this system is clean black and white images with low noise, and the output is images of the separated lines. In this research, we considered a 300 DPI scanned document as input since it is a regular resolution for Persian/Arabic OCR. All of the other internal parameters are fixed for this resolution and for other scanning resolutions, all we need is to scale them proportional to the ratio they have to this predefined DPI. Thus, parameters of the proposed system only depend on the easily defined scanning resolution. Since after scanning a regular size paper such a detailed resolution creates a huge file size, we need to consider some measures to avoid the slow down of the algorithm. Hence, in line segmentation block, most of the time-consuming processes are computed on a downsampled version of the documents image, and only the final processes are performed on the full-size image with some slight adjustment to insure the accuracy.

Figure 5 illustrates the block diagram of the layout analysis subsystem. The downsampling is performed with the ratio of 1/5, so that will result a very rough version of the original image. In Figure 5, H and V smear refer to the process of black painting every neighboring pixels of each back pixel in the boundary of $\pm \Delta$[1]in horizontal or vertical direction respectively.

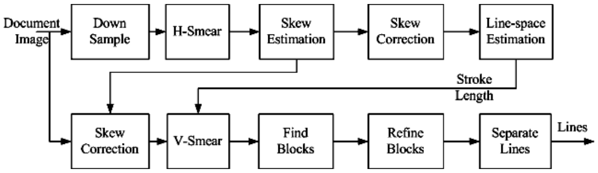
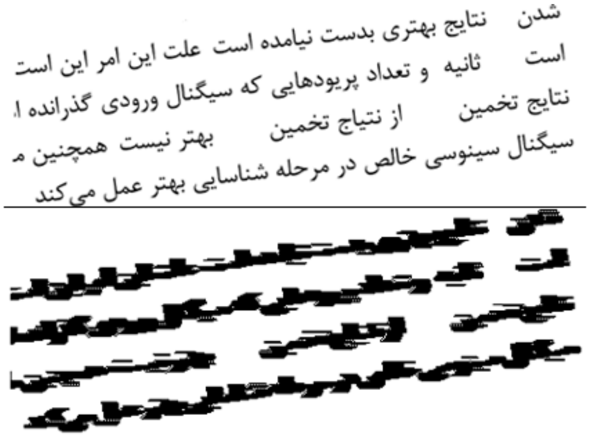


Figure 5. Block diagram of layout analysis subsystem

In the skew detection module, a method similar to [11] was performed to estimate skew angle, and with this estimation, the exhaustive search of skew angle with Hough transform of the full-size image was narrowed down to only a handful of choices. Thus, we guaranteed the required skew detection performance as well as computation speed. In estimation phase, each connected component of the horizontally smeared downsampled image was extracted. The shape of the Persian/Arabic script and their connectivity over the baseline make almost every line to build a single connected component after smearing. Therefore, the

[1] We used $\Delta = 5$ for horizontal smearing.

estimation of the orientation of the connected components yields the estimation of skew degree of lines. Hence, the average orientation of all of the connected components will estimate the expected value of the skew degree in the document (Figure 6).


شدن   نتایج بهتری بدست نیامده است علت این امر این است

است   ثانیه و تعداد پریودهایی که سیگنال ورودی گذرانده ا

نتایج تخمین   از نتیاج تخمین   بهتر نیست همچنین م

سیگنال سینوسی خالص در مرحله شناسایی بهتر عمل می‌کند

Figure 6. Sample of Persian text (Top) and horizontally smeared version of the downsampled image (Bottom)

To calculate the orientation of each component, using least square error method, we computed three straight lines that pass through top, bottom, and middle of it using upper contour, lower contour, and the mean up and lower contour respectively. Thus, averaging the orientation of these lines will yield the estimation of component's orientation (Figure 7).

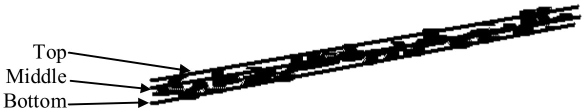


Figure 7. Last line of Figure 6 with the three estimated straight lines at top, bottom, and middle using least square method

The next step toward segmenting the page is to find the text blocks and parsing them into the lines. To make an adaptive algorithm, we design a system to estimate the size of spaces between the lines in the horizontally smeared version of the page using estimation of fundamental frequency in its horizontal projection. For this purpose, the algorithm uses the deskewed version of the horizontally smeared image. Subsequently, it finds the region where the sum of projection is inside the bound defined by h_1 and h_2. These two thresholds are set as 91% and 63% of the maximum value of the projection (Figure 8). Finally, the center of these selected regions are computed using subtractive clustering [12], and average value of distance between the adjacent centers will result the estimation of the space between lines.

This information is used in the full-size image to fill the white spaces between lines with vertical smearing which is entailed by extra dilating operations. Therefore, all lines of every paragraph will be connected with each other, and a search for connected components will result the paragraph blocks. In the refine-block section (Figure 5), an algorithm similar to [13] was performed to connect the very close paragraphs, and separate the paragraphs that have the same inter-space as the lines. The distance between paragraphs were computed as the x-y distance of the enclosing box of the paragraphs. Figure 9 shows the selected paragraphs of a sample page in the vertically smeared and dilated image. Finally, lines of each selected paragraph is simply separated using horizontal projection of the deskewed full-size image.

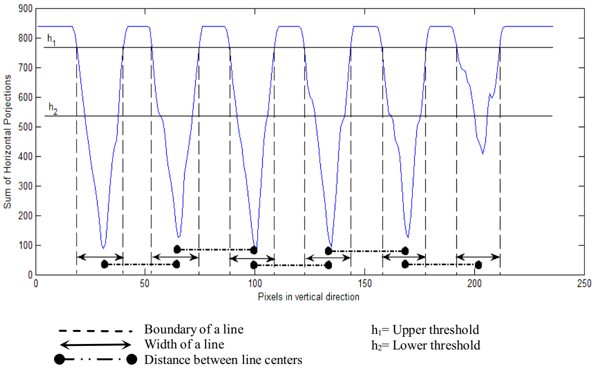
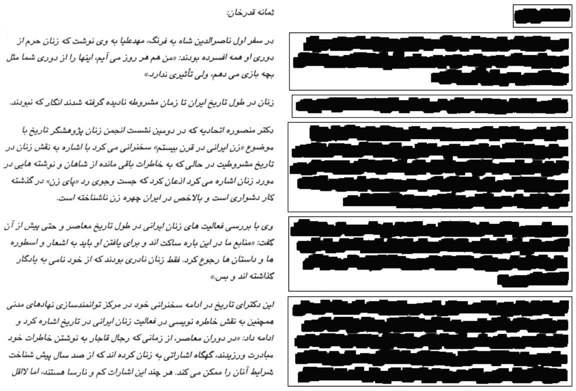


Figure 8. Estimation of spaces between lines and horizontal projection of the horizontally smeared image


ثمانه قدرخان:

در سفر اول ناصرالدین شاه به فرنگ، مهدعلیا به وی نوشت که زنان حرم از دوری او همه افسرده بودند: «من هم هر روز می آیم، اینها را از دوری شما مثل بچه بازی می دهم، ولی تأثیری ندارد.»

زنان در طول تاریخ ایران تا زمان مشروطه نادیده گرفته شدند انگار که نبودند.

دکتر منصوره اتحادیه که در دومین نشست انجمن زنان پژوهشگر تاریخ با موضوع «زن ایرانی در قرن بیستم» سخنرانی می کرد با اشاره به نقش زنان در تاریخ مشروطیت در حالی که به خاطرات باقی مانده از شاهان و نوشته هایی در مورد زنان اشاره می کرد اذعان کرد که جست وجوی رد «پای زن» در گذشته کار دشواری است و بالاخص در ایران چهره زن ناشناخته است.

وی با بررسی فعالیت های زنان ایرانی در طول تاریخ معاصر و حتی پیش از آن گفت: «منابع ما در این باره ساکت اند و برای یافتن او باید به اشعار و اسطوره ها و داستان ها رجوع کرد. فقط زنان نادری بودند که از خود نامی به یادگار گذاشته اند و بس.»

این دکترای تاریخ در ادامه سخنرانی خود در مرکز توانمندسازی نهادهای مدنی همچنین به نقش خاطره نویسی در فعالیت زنان ایرانی در تاریخ اشاره کرد و ادامه داد: «در دوران معاصر، از زمانی که رجال قاجار به نوشتن خاطرات خود مبادرت ورزیدند، گهگاه اشاراتی به زنان کرده اند که از صد سال پیش شناخت شرایط آنان را ممکن می کند. هر چند این اشارات کم و نارسا هستند، اما لااقل

Figure 9. Sample of deskewed page (left) and the corresponding paragraphs in a vertically smeared image. (right)

3.2. Segmentation

The role of segmentation section is to decompose each line image into words, characters, and graphemes. The block diagram of the proposed segmentation method is illustrated in Figure 10. First, PAWs are extracted from the line image using simple search for connected components in the image. Thereafter, some structural features are extracted from each PAW, which will act as the inputs of an ANN. The ANN is a Multi Layer Perceptrons (MLP) that is trained to make high value outputs on the grapheme junction points. Hence, the peaks of ANN output will indicate the most probable points that can be the real junction points. Since we have some obvious rules about wrong places for junction points, a post processing system yields the final junction points.

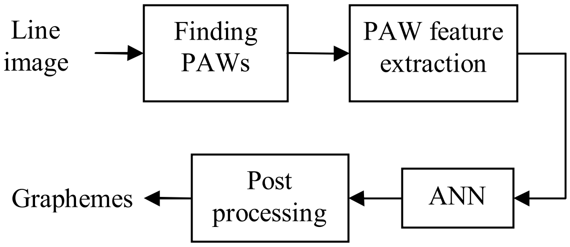


Figure 10. Block diagram of the segmentation module.

To extract the PAW features, some basic characteristics of the script are investigated as the guideline. In Persian/Arabic PAWs, the upper contour has a high gradient in the junction points, and after most junction points, the vertical projection has a value larger than the mean. On the other hand, the pen tip is generally laid near the baseline in the desired junction points.

Considering the above characteristics, three basic features are proposed that together will identify the junction points:

a. Vertical projection of the line image (Figure 11).
b. The first derivative of the upper contour.
c. The distance of the pen tip from the baseline.

See figures 11 and 12, and refer to [14] for more details.

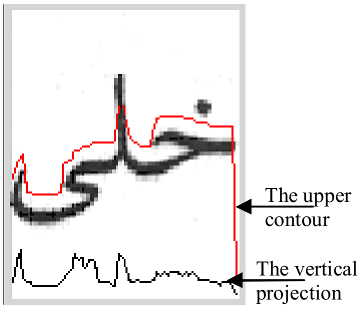


Figure 11. Sample PAW image with contour and vertical projection.

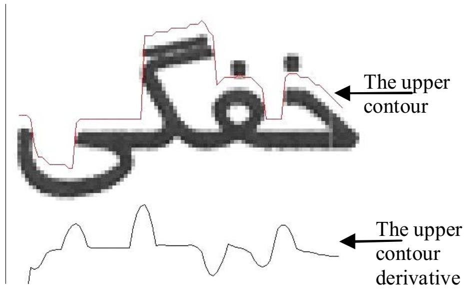


Figure 12. Sample PAW image with contour (red) and its derivative (gray).

In this approach, to get a higher performance, the above features are calculated over a window of width equal to four times the pen width to make a feature vector. The resulting vector is fed to an MLP with one output neuron that estimates the probability of the center point being the junction point.

Our train set includes labeled junction points. The target vector of the neural network should be equal to one for the junction points and zero for the others. To assist the learning procedure, a Gaussian function with variance equal to 1/6 times the pen width is placed at each junction point. Although this smoothing reduces the accuracy, the results are quite acceptable (Figure 13).

In practice, our neural network should have a predefined number of inputs; consequently, the input image is normalized with the pen width value in order to have pen width equal to five points in all images. In the implementation stage, the neural network window width is set to 20, i.e. four times the pen width. Thus, the neural network has 60 neurons in the input layer and five hidden neurons.

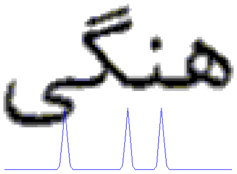


Figure 13. Sample PAW image neural network target signal which is used for training procedure

The described neural network uses a hyperbolic tangent sigmoid as the transfer function in all neurons, and it is trained using the standard gradient decent with momentum training algorithm which has adaptive learning rate [15]. In order to facilitate the training process and escape from the local minimums, a simulated annealing algorithm is added to the training method. The peaks of the neural network output are selected as the candidate junction points. Finally, the neighboring points of the candidates are checked for the location of the pen tip, and the final junction points are specified

3.3. Recognition

The recognition subsection will translate the input grapheme image into a single character or a symbol. In Persian script, every letter can have two or four different shapes in respect to their position in their containing PAW. The four different positions are at the beginning, in the middle, at the end, or isolated word. These positions correspond to the connectivity of the letter from left, both sides, right, or no connection respectively. Table 1 shows the example of four different shapes of character "HEH". The explained arrangement of the characters is addressed in previous works [7, 16] as the main characteristic of Persian/Arabic cursive scripts. This feature will increases the literally available 52 characters of Persian script to 134 different characters in shape. To overcome this diversity, four different recognition systems will be generated for each position of characters.

In the proposed algorithm, a group number is assigned to each case of character shapes (Table 1). In contrast to the methods available in the literature, instead of applying classifiers to recognize characters, the classifiers recognize graphemes in the four positions inside words. Replacing the character recognition with grapheme recognition, will reduce the number of recognition classes from 134 characters to 85 graphemes. Meanwhile, a post-processing system will combine the information of number and position of dots, and the sequence of the graphemes to produce the final recognized characters. This novel policy will narrow down the complexity of the segmentation and classification sections significantly.

Table 1. Four possible shapes of a character in a PAW and corresponding character group of each shape and its connectivity direction.

Group Number	*1*	*2*	*3*	*4*
Character "HEH"	ﻫ	ﻬ	ﻪ	ه
Connectivity Direction	Left	Both	Right	None

For every grapheme image, a feature vector of length 50 is computed which consists of normalized values of both statistical and structural features. The former is mainly gathered from the statistical distribution of the grapheme skeleton [16], while the latter is mostly related to the shape and morphological characteristics of the Persian script [17, 18]. We encourage the interested readers to find the detailed information about the selected features in [14] as we avoid restating them here.

The SVM classifier with RBF kernel was used to recognize the graphemes, and Table 2 illustrates the selected parameters which are optimized by cross-validation.

Table 2. SVM Optimized Parameters

Group	Kernel	γ	c	SVs	Classes
1	RBF	0.01	1	1195	13
2	RBF	0.01	2.2	951	12
3	RBF	0.01	2.2	930	20
4	RBF	0.01	1	1664	40

4. Implementation and results

In order to have confidence in the results, a comprehensive database of characters and documents is needed and since there was no standard dataset available for Persian script, we decided to build it from scratch. We gathered various documents in single and double column layouts for each font type and size without any graphical elements. These documents were printed and scanned using desktop equipments.

Additionally, since the proposed algorithm exploits learning systems in segmentation section as well as in the recognition subsystem, two more datasets are needed. The first set should contain data samples of the labeled PAWs to train and test the segmentation ANN, and the second set should have a complete set of labeled graphemes to train and test the SVM classifier.

To build a semi-automatic mechanism for creating the first set, a primitive segmentation system was designed according to some trivial If-Then rules that perform over the PAW images. Thereafter, this system was used to segment about 40 pages of Persian script from daily newspapers in different font types, and its results have been verified manually. As the result, the labeled dataset has been created which was partitioned into with ratio of 1 to 3 for test and train purposes.

On the way to have complete datasets, we trained the segmentation neural network with the above set, and used this system to segment a large number of printed documents in 20 fonts to create four groups of grapheme database. Figure 14 shows an example of some of these fonts. This dataset is also verified to have a complete multi-font labeled dataset of Persian printed documents. Our database compromises over 40,000 sample PAWs for segmentation and 175,632 graphemes. The train and test set are chosen uniformly random with ratio of 1/3.

MATLAB Neural Network toolbox has been used to produce the prototype version of algorithm along with the help of OSU-SVM toolbox. Employing Delphi platform as the final implementation environment has increased the speed and helped the efficiency of the algorithm.

Table 3 illustrates the average results of four different page skews. The observations show that the proposed algorithm has slight lower accuracy for the pages with smaller skew degree. We interpreted this as the result of more distortion occurred in the interpolation of the low-angle rotated images. Tables 4 to 6 provide the detailed results of our system for segmentation, recognition, and overall results.

5. Conclusion

With the proposed design, we have achieved an accurate OCR which is independent of font size for Persian/Arabic printed documents with ability to recognize omni-font scripts. Moreover, this system segments the full-text Manhattan style documents free of font size or page layout size using a single adjustable parameter, and it perfectly tolerates the skew of 20 degrees in the scanned pages. The segmentation and recognition sections of the proposed system use learning systems, which ensures adaptiveness and flexibility of the algorithm. The commercial version of this system has been applied in Iranian Civil Organization registration in year 2005.

6. Acknowledgement

The authors would like to thank and acknowledge the Paya Soft co. that sponsored this research.

Table 3. Recognition rate for different page skews on a randomly selected 4 pages from database

Skew (degrees)	NLP	AWPP[2]	AFWPP[3]	WRR[4]
5	Yes	438	19	95.66 %
10	Yes	438	16	96.35 %
16	Yes	438	15	96.57 %
20	Yes	438	17	96.12 %

Table 4. Correct Segmentation Rate

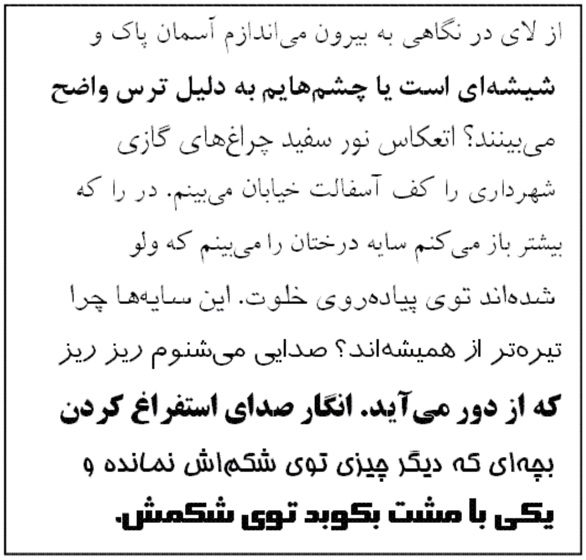
Segmentation	Train set	Test Set
ANN	99.4 %	98.7 %

Table 5. Grapheme Classification Rate

Group	Samples	Train Set	Test Set
1	44672	99.6 %	99.2 %
2	40885	99.8 %	98.8 %
3	40123	99.8 %	98.6 %
4	49952	99.1 %	96.3 %

Table 6. Overall Recognition Rate

Type	Train Set After NLP	Test Set
Character	-	98.3 %
Word	94.19 %	90.17 %



از لای در نگاهی به بیرون می‌اندازم آسمان پاک و
شیشه‌ای است یا چشم‌هایم به دلیل ترس واضح
می‌بینند؟ اتعکاس نور سفید چراغ‌های گازی
شهرداری را کف آسفالت خیابان می‌بینم. در را که
بیشتر باز می‌کنم سایه درختان را می‌بینم که ولو
شده‌اند توی پیاده‌روی خلوت. این سایه‌ها چرا
تیره‌تر از همیشه‌اند؟ صدایی می‌شنوم ریز ریز
که از دور می‌آید. انگار صدای استفراغ کردن
بچه‌ای که دیگر چیزی توی شکم‌اش نمانده و
یکی با مشت بکوبد توی شکمش.

Figure 14. Example of a Persian script with different font types in each line.

[2] Average Word per Page

[3] Average Faulty Word per Page

[4] Word Recognition Rate

0-7695-2467-2/05 $20.00 © 2005 IEEE

7. 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, Nov. 1998.

[3] G. NAGY, "Twenty years of document image analysis in PAMI", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, 1, pp. 38-62, Jan 2000.

[4] B. Al-Badr, RM. 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., Los Alamitos, CA, USA, pp. 355-359, 1995.

[5] 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, June 1999.

[6] AH. 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, Jul 2004.

[7] A. Cheung, M. Bennamoun, NW. Bergmann, "An Arabic optical character recognition system using recognition-based segmentation", *Pattern Recognition*, vol. 34, no. 2, Elsevier, pp. 215-233, UK., Feb. 2001.

[8] H. Weissman,M. Schenkel,I. Guyon,C. Nohl,D. Henderson, "Recognition-based segmentation of on-line run-on handprinted words: input vs. output segmentation", *Pattern Recognition*, vol. 27, no. 3, UK., pp. 405-420, March 1994.

[9] R. Azmi, E. Kabir, "A new segmentation technique for ominifont farsi text", *Pattern Recognition Letters*, vol. 22, no. 2, pp. 97-104, 2001.

[10] R.C. Gonzalez, P. Wintz, *Digital Image Processing*, Addison Wesley Publishing Company, 2nd Edition, pp. 392-423, 1987.

[11] Zhixin Shi, Venu Govindaraju, "Skew Detection for Complex Document Images Using Fuzzy Runlength", ICDAR 2003: pp. 715-719, 2003.

[12] S.L. Chiu, "Fuzzy model identification based on cluster estimation", *Journal of Intelligent Fuzzy Systems*, vol. 2, pp. 267-278, 1994.

[13] A. Simon, J.C. Pret, and A.P. Johnson, A fast algorithm for bottom-up document layout analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, 3 (March 1997), pp. 273-277.

[14] H. Pirsiavash, R. Mehran, and F. Razzazi, "A Robust Free Size OCR for Omni-font Persian/Arabic Printed Document using MLP/SVM," 10th Iberoamerican Congress on Pattern Recognition, Havana, Cuba, Springer LNCS 3773, pp. 601–610, Nov. 2005.

[15] S. Haykin, *Adaptive Filter Theory*, 3rd edition, Upper Saddle River, NJ: Prentice-Hall, 1996.

[16] H. Pirsiavash F. Razzazi, "Design and Implementation of a Hierarchical Classifier for Isolated Handwritten Persian/Arabic Characters", IJCI Proceedings of International Conference on Signal Processing, Vol. 1, no. 2, Turkey, September 2003.

[17] M. Kavianifar, A. Amin, "Preprocessing and structural feature extraction for a multi-fonts Arabic/Persian OCR", Conference on Document Analysis and Recognition, IEEE Computer Soc., pp. 213-216. Los Alamitos, CA, USA, 1999.

[18] BM. Kurdy, MM. Al Sabbagh, "Omnifont Arabic optical character recognition system", Proceedings of Int. Conf. on Information and Communication Technologies: From Theory to Applications, IEEE, Piscataway, NJ, USA, pp. 469-70, 2004.