FHT: An Unconstraint Farsi Handwritten Text Database

Majid Ziaratban [a] Karim Faez [a] Fatemeh Bagheri [b]
[a] *Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran*
[b] *Computer Engineering and IT Department, Amirkabir University of Technology, Tehran, Iran*
{m_ziaratban , kfaez} @aut.ac.ir

Abstract

Standard databases play very important roles in pattern recognition tasks. To compare the performances of different algorithms, they must be tested on a same dataset. In Farsi, there is not a database of handwritten texts to evaluate different algorithms. In this paper, an unconstraint Farsi handwritten text database is introduced. 250 participants in different ages and education levels filled 1000 forms. Duo to the characteristics of the database, it can be used in many OCR applications. A large number of writers, big lexicon size, out-of-straight textlines and various categories of texts are some of the characteristics.

1. Introduction

The optical character recognition was started from the recognition of machine printed digits and characters. It was developed to the recognition of machine printed words. Gradually, handwritten digit, character and word recognition were introduced into this domain. Most researches have been done in Latin languages. Thus, the recognition task has been promoted rather than the other languages, so that, nowadays the situations such as English sentence recognition and understanding using grammars and natural linguistic syntax are under study and some researches [1] have been performed in this field. While in some other languages, the growth has not been developed as fast as in Latin. Farsi and Arabic are two similar languages in alphabets in which many challenges are still remained in the word recognition domain.

Great similarities among different characters, wide varieties in writing styles, different shapes for a character with respect to its location in a word, dots and diacritics, and the cursive writing are some important factors which cause the Farsi/Arabic word recognition to be very difficult. In these languages, in last years, most researchers have been focused on solving the problems of the handwritten word recognition.

The next step in Farsi and Arabic OCR is the recognition of the words in a text and then understanding the sentences. However, some sparse researches have been done and started this field. They studied on the recognition of bank check amounts [2,3,4]. They can be preamble of using grammars and recognizing texts in Farsi and Arabic scripts.

One of the reasons of the slower development of Farsi OCR rather than the Latin languages is the intrinsic features of the Farsi scripts. Another substantial reason is the lack of standard databases of Farsi digits, characters, words and texts. Standard databases play vital roles in pattern recognition tasks. To compare different algorithms and select the best ones, they must be tested on a same dataset. Only the results obtained from standard databases can be reliable and used for evaluating the performances of various approaches. Consequently, standard databases can strongly improve the OCR researches.

Many databases in the handwritten recognition domain have been gathered and used in various languages and applications. Some widely used databases are NIST [5], CENPARMI [6], CEDAR [7], UNIPEN [8], ETL9 [9] and PE92 [10]. There are databases in Latin [5-7, 11-14], Chinese [15], Korean [10], Indian [16], Arabic [3,4, 17-20] and Farsi [2,21,22] for offline handwritten recognition applications. Many of them are consisted of handwritten isolated digits [2,3,7,12,13,17,19,20, 23], characters [7,10,12,13,19,20,23] or words [3,4,7, 13, 17-19, 21-23] and few of them include sentences or texts [11,14,15]. Some information about related databases is given in Table 1. As we can see in this table, there is not any database of handwritten texts in Farsi.

In this paper, a Farsi handwritten text database, *FHT*, is introduced which is useful for many OCR applications. The rest of the paper is organized as follows: Section 2 describes an overview of the FHT. Section 3 and 4 respectively discuss about the layout design and some statistics of the database. In section 5, the ground truth is described. Finally, conclusions are drawn in section 6.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.56

IEEE computer society

Table 1. Related databases in different languages

Ref.	Database name	Year	Language	Content	Lexicon size	Number of writers	Number of writers per form	
[10]	PE92	1993	Korean	Isolated characters	-			235000 characters
[7]	CEDAR	1994	English	Single words (city names), Single characters and digits				5000 words 50000 characters and digits
[11]	CAMBRIDGE	1998	English	**Sentences**	1334	1	1	4051 words
[17]	-	1999	Arabic	Single words, digits and signatures				
[12]	GRUHD	2001	Greek	A small text (poem), characters, digits and symbols		1000		1760 forms 102692 words 123256 digits
[13]	-	2002	Italian	Single words (used in Italian legal check amounts), single characters, digits and signatures	49	277		28678 digits 66609 characters 48584 words 2222 signatures
[14]	IAM	2002	English	**Sentences**	10841	400		1066 forms 82227 words
[18]	IFN/ENIT	2002	Arabic	Single words (Tunisian city names)	946	411	5	2265 forms 26459 words
[3]	-	2003	Arabic	Legal and courtesy check amounts				2499 amounts 29498 subwords 10425 digits
[4]	AHDB	2004	Arabic	Legal check amounts	96	100		105 forms
[15]	HIT-MW	2007	Chinese	**Sentences**	3041	780		853 forms 8664 textlines 186444 characters
[2]	-	2007	Farsi	Legal and courtesy check amounts	40	100	1	100 forms 8400 words 14600 subwords 5900 digits
[21]	IfN/Farsi	2008	Farsi	Single words (Iranian city names)	1080	600		7271 word images 23545 subwords
[22]	IAUT/PHCN	2008	Farsi	Single words (Iranian city names)	1140	380	3	1140 forms 34200 word images 107310 subwords
[23]	-	2008	Dari	Single words, single characters and digits	73	200		28000 digits 7400 characters 14600 words
[19]	-	2008	Arabic	Single words, single characters and digits	70	328		13439 digits 21426 characters 11375 words
[20]	LMCA	2008	Arabic	Single words, single characters and digits		55		30000 digits 100000 characters 500 words

2. Overview of the FHT

In order to collect a database with naturally written texts, forms were consigned to participants to fill it with enough time and without any stress. The texts were printed with a large enough font to avoid writers from writing more carefully than their daily writing.

Unlike IAM, participants did not use any rulers to adjust their writing paths. Thus the database can be used for textline extraction and baseline correction researches.

Participants could write by any writing instruments and no restrictions were imposed on them. The filled forms have been scanned at 300 dpi in 256 gray scales.

The structure and foundation of our database is inspired by the IAM database which is the standard collection of the handwritten English full sentences.

In our dataset, the texts were sampled from corpus in different categories. These categories were selected similar to the IAM database and are given in Table 2 with the number of filled forms in each category.

Duo to the characteristics of the FHT database, it can be used in many OCR applications as follows:

1- Word and subword recognition
2- Segmentation the words into characters
3- Baseline detection and textline extraction
4- Discrimination between machine printed and handwritten texts

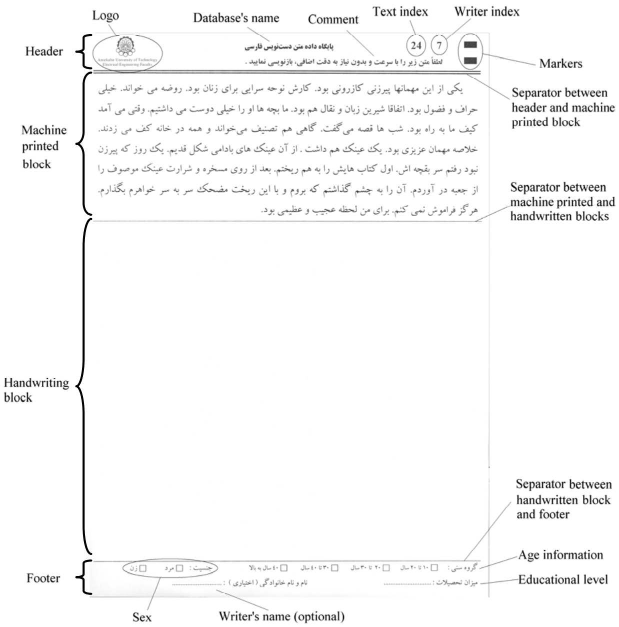
5- Lexicon reduction
6- Writer identification
7- Layout analysis
8- Document classification
9- Farsi sentence recognition and understanding

3. Layout design

The layout of forms must be simple and clear for writers. Also it must be simple for automatically reading by a form reader system. Two markers were designed to determine the top-right coordinate of the forms. The layout is divided into four distinct blocks: header, machine printed block, handwriting block and footer. The horizontal lines separate adjacent blocks. Some information about the writer is taken in the footer block. More details are illustrated in Figure 1. A sample filled form is depicted in Figure 2.

Table 2. The categories of texts

	Subject	Number of filled forms
1	Press: sports	50
2	Press: articles	175
3	Press: economic	100
4	Press: cinema	25
5	Religion	150
6	Fictions	50
7	Imaginary fiction	50
8	Biography	50
9	Poem	50
10	Literature	125
11	Scientific articles	25
12	Check amounts	25
13	Miscellaneous	125


Logo
Database's name
Comment
Text index
Writer index
Header
Markers
Separator between header and machine printed block
Machine printed block
Separator between machine printed and handwritten blocks
Handwriting block
Separator between handwritten block and footer
Age information
Educational level
Footer
Sex
Writer's name (optional)
Amirkabir University of Technology
Electrical Engineering Faculty
24
7

پایگاه داده متن دست‌نویس فارسی
لطفاً متن زیر را با سرعت و بدون نیاز به دقت اضافی، بازنویسی نمایید .

یکی از این مهمانها پیرزنی کازرونی بود. کارش نوحه سرایی برای زنان بود. روضه می خواند. خیلی
حراف و فضول بود. اتفاقا شیرین زبان و نقال هم بود. ما بچه ها او را خیلی دوست می داشتیم. وقتی می آمد
کیف ما به راه بود. شب ها قصه می‌گفت. گاهی هم تصنیف می‌خواند و همه در خانه کف می زدند.
خلاصه مهمان عزیزی بود. یک عینک هم داشت . از آن عینک های بادامی شکل قدیم. یک روز که پیرزن
نبود رفتم سر بقچه اش. اول کتاب هایش را به هم ریختم. بعد از روی مسخره و شرارت عینک موصوف را
از جعبه در آوردم. آن را به چشم گذاشتم که بروم و با این ریخت مضحک سر به سر خواهرم بگذارم.
هرگز فراموش نمی کنم. برای من لحظه عجیب و عظیمی بود.

گروه سنی : □ ۱۰ تا ۲۰ سال □ ۲۰ تا ۳۰ سال □ ۳۰ تا ۴۰ سال □ ۴۰ سال به بالا جنسیت : □ مرد □ زن
میزان تحصیلات : .... نام و نام خانوادگی (اختیاری) : ....

Figure 1. Form layout

4. Statistics

The forms were filled by 250 writers in different ages and education levels. 65% of them were male and the rests were female.

Each participant was asked to write four forms located in a same category. A text in a category was written by 25 different writers.

The database contains 1000 filled forms. It includes totally 106600 handwritten Farsi words and 230175 subwords distributed in 8050 sentences.

In average, each filled form comprises 6.45 textlines, 8.05 sentences, 106.6 words, 230.175 subwords, 406 characters and 132.1 dots. Each textline contains 16.53 words and each word includes 2.16 subwords and 3.81 characters. Also 13.24 words are made a sentence, in average.

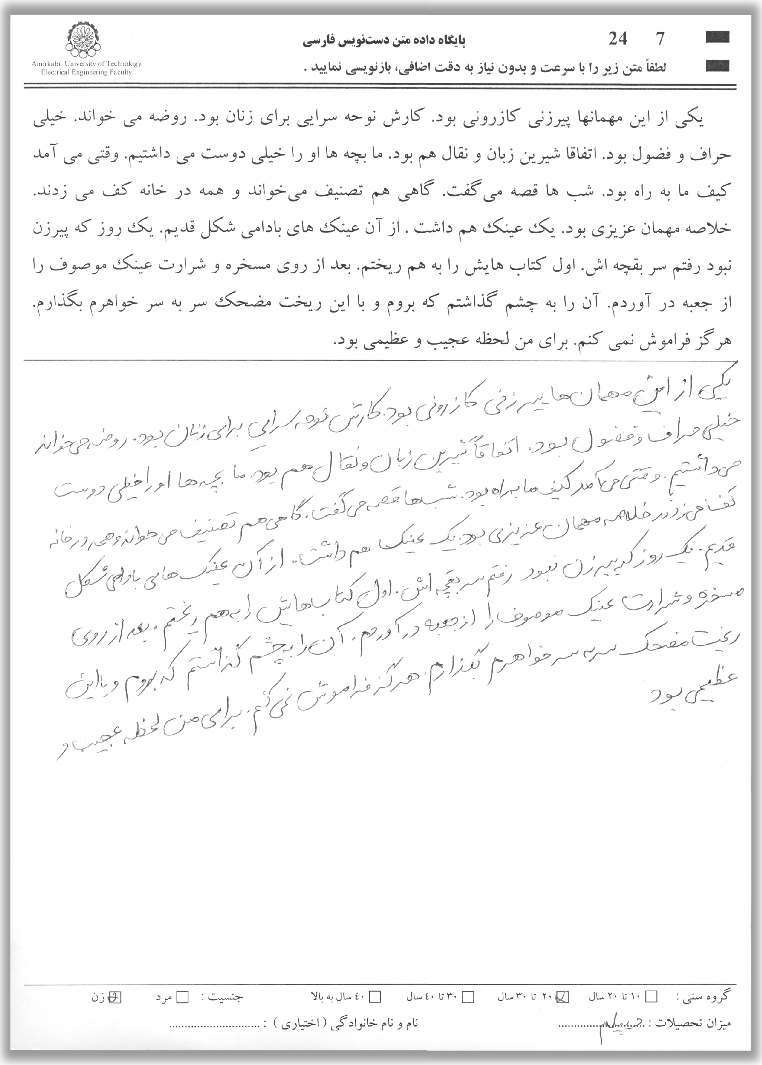
A great number of dots in a text is one of the characteristics of Farsi/Arabic scripts rather than other languages. In the database, there are 1.24 dots per word.

Subwords contained one to six characters i.e. there was not any subword with more than 6 characters. The distribution of the subwords is shown in Figure 3. The lexicon consists of 1410 different words and 724 dissimilar subwords. The distribution of the dissimilar subwords versus the number of characters in a subword is given in Figure 4. It can be observed that the subwords with only one character construct about half of all subwords in Figure 3, while their variety is low (See Figure 4). It means that there are many samples for each of dissimilar one-character subwords. Figure 5 shows the average number of samples for each dissimilar subword regarding to the number of characters in a subword.

5. Ground truth

Making *ground truth* (GT) files is a very important part of a database. This is indeed a prerequisite for any processes on the database, particularly the recognition tasks. To compute recognition rates, some information about the written texts in forms are needed.

On one hand, the accuracy of the labeling is very important. On the other hand, it is an error-prone and time consuming task. To create a primary version of GT files, first, the machine printed text of a form is completely copied into a GT file. Then, the written texts in the filled forms were manually inspected three times to ensure about the correspondence between the handwritten text and the text in its related GT file. Handwritten texts almost have been written completely and thoroughly; but in some cases, there were some differences. Thus some corrections should be done.



Amirkabir University of Technology
Electrical Engineering Faculty

24 7

پایگاه داده متن دست‌نویس فارسی

لطفاً متن زیر را با سرعت و بدون نیاز به دقت اضافی، بازنویسی نمایید .

یکی از این مهمانها پیرزنی کازرونی بود. کارش نوحه سرایی برای زنان بود. روضه می خواند. خیلی
حراف و فضول بود. اتفاقا شیرین زبان و نقال هم بود. ما بچه ها او را خیلی دوست می داشتیم. وقتی می آمد
کیف ما به راه بود. شب ها قصه می گفت. گاهی هم تصنیف می‌خواند و همه در خانه کف می زدند.
خلاصه مهمان عزیزی بود. یک عینک هم داشت . از آن عینک های بادامی شکل قدیم. یک روز که پیرزن
نبود رفتم سر بقچه اش. اول کتاب هایش را به هم ریختم. بعد از روی مسخره و شرارت عینک موصوف را
از جعبه در آوردم. آن را به چشم گذاشتم که بروم و با این ریخت مضحک سر به سر خواهرم بگذارم.
هر گز فراموش نمی کنم. برای من لحظه عجیب و عظیمی بود.

یکی از این مهمان ها پیرزنی کازرونی بود. کارش نوحه سرایی برای زنان بود. روضه می خواند
خیلی حراف و فضول بود. اتفاقا شیرین زبان و نقال هم بود. ما بچه ها او را خیلی دوست
می داشتیم. وقتی می آمد کیف ما به راه بود. شب ها قصه می گفت. گاهی هم تصنیف می خواند و همه در خانه
کف می زدند خلاصه مهمان عزیزی بود. یک عینک هم داشت. از آن عینک های بادامی شکل
قدیم. یک روز که پیرزن نبود رفتم سر بقچه اش. اول کتاب هایش را به هم ریختم. بعد از روی
مسخره و شرارت عینک موصوف را از جعبه در آوردم. آن را به چشم گذاشتم که بروم و با این
ریخت مضحک سر به سر خواهرم بگذارم. هرگز فراموش نمی کنم. برای من لحظه عجیب و
عظیمی بود

گروه سنی : □ ۱۰ تا ۲۰ سال ☑ ۲۰ تا ۳۰ سال □ ۳۰ تا ۴۰ سال □ ۴۰ سال به بالا جنسیت : □ مرد ☒ زن

میزان تحصیلات : دیپلم .... نام و نام خانوادگی (اختیاری) : ....

Figure 2. A sample filled form

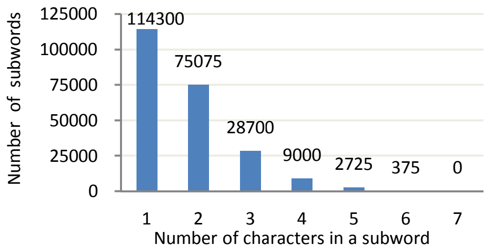


Figure 3. The distribution of subwords versus the number of characters in a subword

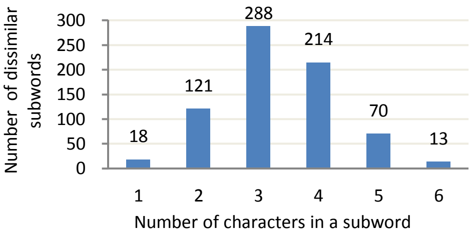


Figure 4. Variety of subwords

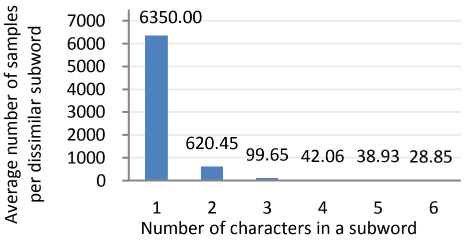
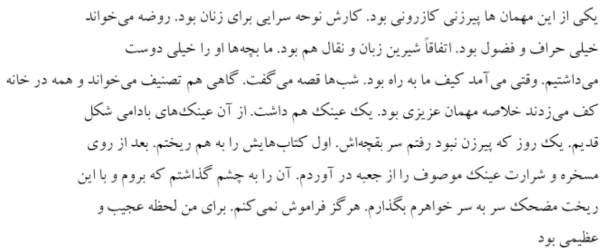


Figure 5. The average number of samples for each dissimilar subword regarding to the number of characters in a subword

Omission and repetition of one or more words are the main reasons of the inequalities. They are corrected with deletion and insertion the mentioned words respectively from and into the GT file. Also in some forms, a word has been written different from the original one in the printed text. In these cases, the corresponding word in the GT file is changed into the word written in the form.

Another type of corrections in GT files is about textlines. The font size of the printed texts in the forms has been selected large enough, so that a word in a printed text, approximately occupies an equal space in a handwritten text. Hence, writers could write each printed textline, completely in one separated textline. Because of wide varieties in sizes of handwritten words and the gaps between two succeeding words written by different writers, in many cases, the correspondence between printed and handwritten textlines have not been respected. Hence, a textline matching process is needed to obtain the GT files as similar as possible to the handwritten texts.

The ground truth of the sample form in Figure 2 is shown in Figure 6.



یکی از این مهمان ها پیرزنی کازرونی بود. کارش نوحه سرایی برای زنان بود. روضه می‌خواند
خیلی حراف و فضول بود. اتفاقاً شیرین زبان و نقال هم بود. ما بچه‌ها او را خیلی دوست
می‌داشتیم. وقتی می‌آمد کیف ما به راه بود. شب‌ها قصه می‌گفت. گاهی هم تصنیف می‌خواند و همه در خانه
کف می‌زدند خلاصه مهمان عزیزی بود. یک عینک هم داشت. از آن عینک‌های بادامی شکل
قدیم. یک روز که پیرزن نبود رفتم سر بقچه‌اش. اول کتاب‌هایش را به هم ریختم. بعد از روی
مسخره و شرارت عینک موصوف را از جعبه در آوردم. آن را به چشم گذاشتم که بروم و با این
ریخت مضحک سر به سر خواهرم بگذارم. هرگز فراموش نمی‌کنم. برای من لحظه عجیب و
عظیمی بود

Figure 6. The ground truth correspond to the handwritten text in Figure 2

6. Conclusions

A database of Farsi handwritten texts and sentences was introduced in this paper. This is the only dataset in Farsi which contains handwritten texts. The lexicon used in the dataset is bigger than lexicons of current Farsi word datasets. Hence, the lexicon reduction approaches can use it to compare their performances. The structure and foundation of our database is inspired by the IAM database; but unlike the IAM, writers did not use any rulers to write the texts. Thus, our dataset can also be used in textline extraction applications. 250 participates wrote texts. Each text was written by 25 writers. Hence, the FHT database can be useful in writer identification. The database can be used in many other applications such as word/subword recognition, segmentation the words into characters, discrimination between machine printed and handwritten texts, layout analysis, document classification, and Farsi sentence recognition and understanding. FHT is available for academic researches by contacting with the authors.

7. References

[1] R. Bertolami and H. Bunke, "Hidden Markov model based ensemble methods for offline handwritten text line recognition", *Pattern Recognition* 41(11), 2008, pp. 3452-3460.

[2] M Ziaratban, K Faez and M Ezoji, "Use of Legal Amount to Confirm or Correct the Courtesy Amount on Farsi Bank Checks". *ICDAR*, 2007, pp. 1123-1127.

[3] Y. Al-Ohali, M. Cheriet and C.Y. Suen, "Databases for recognition of handwritten Arabic cheques". *Pattern Recognition.* 36, 2003, pp. 111-121.

[4] S. Alma'adeed, D. Elliman, and C.A. Higgins*, "*A Data Base for Arabic Handwritten Text Recognition Research*," Int. Arab J. Inf. Technol.* 1(1), 2004.

[5] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, C. Wilson, The first census optical character recognition systems conf. #NISTIR 4912, The U.S. Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD, 1992.

[6] C.Y. Suen, C. Nadal, R. Legault, T. Mai, L. Lam, "Computer recognition of unconstrained handwritten numerals," *Proc. of the IEEE*, 7(80), 1992, pp.1162–1180.

[7] J. Hull, "A database for handwritten text recognition research," *IEEE Trans. on PAMI*, 16(5), 1994, pp.550–554.

[8] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, S. Janet, "Unipen project of on-line data exchange and benchmarks," *ICPR*, 1994, pp. 29–33.

[9] T. Saito, H. Yamada, K. Yamamoto, "On the data base ETL 9 of handprinted characters in JIS Chinese characters and its analysis," *IEICE Transactions,* J68-D(4), 1985, pp.757–764.

[10] D. Kim, Y. Hwang, S. Park, E. Kim, S. Paek, S. Bang, "Handwritten korean character image database PE92," *ICDAR*, 1993, pp. 470–473.

[11] A.W. Senior, A.J. Robinson, "An off-line cursive handwriting recognition system", *IEEE Trans. on PAMI* 20(3), 1998, pp. 309–321.

[12] E. Kavallieratou, N. Liolios, E. Koutsogeorgos, N. Fakotakis, G. Kokkinakis, "The GRUHD Database of Greek Unconstrained Handwriting". *ICDAR*, 2001, pp. 561-565

[13] G. Dimauro, S. Impedovo, R. Modugno, G. Pirlo, "A New Database for Research on Bank-check Processing", *IWFHR*, 2002, pp. 524-528.

[14] U.V. Marti, H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *IJDAR* 5(1), 2002, pp. 39-46.

[15] T.-H. Su, T.-W. Zhang and D.J. Guan, "Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text"*, IJDAR* 10(1), 2007, pp. 27-38.

[16] U. Bhattacharya, B.B. Chaudhuri, "Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals", *IEEE Trans. on PAMI* 31, 2009.

[17] N. Kharma, M. Ahmed, R. Ward, "A New Comprehensive Database of Hand-written Arabic Words, Numbers, and Signatures used for OCR Testing", *IEEE Canadian Conference on Electrical & Computer Engineering*, 1999, pp. 766-799

[18] M. Pechwitz, S.S. Maddouri, V. Maergner, N. Ellouze, H. Amiri "IFN/ENIT - database of handwritten Arabic words", *CIFED'02*, 2002 , pp. 129-136

[19] H. Alamri, J. Sadri, C.Y. Suen, N. Nobile, "A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition", *ICFHR*, 2008.

[20] M. Kherallah, A. Elbaati, H. E. Abed, A.M. Alimi, "The On/Off (LMCA) Dual Arabic Handwriting Database", *ICFHR*, 2008.

[21] S. Mozaffari, H.E. Abed, V. Margner, K. Faez, A. Amirshahi, "IfN/Farsi-Database: A Database of Farsi Handwritten City Names", *ICFHR*, 2008.

[22] A.M. Bidgoli, M. Sarhadi, "IAUT/PHCN : Azad University of Tehran / Persian Handwritten City Names, a very large database of handwritten Persian word", *ICFHR*, 2008.

[23] M.I. Shah, J. Sadri, C.Y. Suen, N. Nobile, "A New Multipurpose Comprehensive Database for Handwritten Dari Recognition", *ICFHR*, 2008.