A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition

Huda Alamri *Javad Sadri* *Ching Y. Suen* *Nicola Nobile*

CENPARMI (Center for Pattern Recognition and Machine Intelligence), Computer Science and Software Engineering Department, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, Quebec, Canada, H3G 1M8, Tel: (514)-848-2424-Ext: 7950, Fax :(514)-8482830.
Emails: {hu_alam, j_sadri, suen, nicola}@cse.concordia.ca

Abstract

This paper presents the work toward developing a new comprehensive database for Arabic off-line handwriting recognition. The database includes: isolated Indian digits, numerical strings, Arabic isolated letters,[i] and a collection of 70 Arabic words. Also, the database includes a free format sample of an Arabic date. A data entry form was designed to collect written samples from Arabic native speakers. Our database is advanced in terms of the variety of sets, words and number of the participants involved. The databases have been divided into respective training, testing and validation sets which will be available in the future for the handwriting recognition community.

Keywords: Arabic Handwritten Recognition, Arabic OCR, Handwritten Segmentation, Farsi handwritten recognition.

1. Introduction

One of the most challenging aspects of off-line handwriting recognition is finding a good database that well represents the variety of handwriting styles and contains the most important classes in the target language. To the best of our knowledge, no such database for Arabic handwriting which contains all the different classes of the language, such as isolated digits, numeral strings, isolated characters, and words is available.

Arabic language is a member of the semantic languages. It is ranked as the sixth most widely spoken language in the world [1], spoken by more than 256 million people in over 20 countries such as Algeria, Bahrain, The Comoros, Chad, Egypt, Eritrea, Iraq, Jordan, Lebanon, Libya, Mauritania, Morocco, Oman, Saudi Arabia, Syria, Tunisia, United Arab Emirates (UAE), and Yemen [2].

Arabic words are written in a cursive-script manner from right to left, where the letters in each word are connected in a certain manner. Therefore, the shape of a letter may change significantly depending on its position within a word [3].

Besides Arabic letters, Indian digits are used in the Arabic scripts, where Arabic digits are used in Latin scripts. As in Latin, numerical strings in Arabic are written from the left to right.

This paper is organized as follows: The next section describes related works toward developing different Arabic off-line handwriting databases. Then, the data collecting stage is presented in Section 3. After that, data extraction and the pre-processing methods are described in Section 4. In Section 5, the database structure and its different datasets are presented. In Section 6, ground-truth data is presented. The results of the segmentation are represented in Section 7. Finally, we present conclusion and future work.

2. Related works

In the last ten years, a lot of research has been devoted to development of databases for handwritten Latin-scripts [4] [5] [6]. However, there has not been much effort toward developing comprehensive databases for Arabic handwriting recognition [3].

For Arabic handwritten word recognition, IFN/ENIT database was developed in 2002, by the Institute of Communications Technology (IFN) at Technical University Braunschweing in Germany and the The National School of Engineers of Tunis (ENIT). It consists of 26,549 images of Tunisian town/village names written by 411 writers [6]. Another Arabic database for handwritten words is the AHDB database, developed in 2003 by Alma'adeed [8]. It includes images of words that are used to describe numbers and quantities in checks, images of the most frequent words used in Arabic writing, and images of sentences used in writing legal amount on Arabic checks. In 2003, Al-Ohali et al. [9], of the Center for Pattern Recognition and Machine Intelligence (CENPARMI), developed an Arabic check database for research in recognition of Arabic handwritten checks. The database includes images for Arabic legal amounts, and

Arabic sub-words (mainly used in writing legal amounts, courtesy amounts, and Indian digits).

3. Data Collection

Finding a real source to collect samples for the target Arabic sets could be quite difficult and could involve a lot of complicated pre-processing tasks, such as removing noises, unwanted data and a lot of segmentation processes. Therefore, the ideal solution was to design a specific data-entry form and collect samples for those data sets from Arabic native speakers. We designed a form consisting of two pages. The first page includes: a sample of an Arabic date, 20 isolated digits as 2 samples of each isolated digit, 38 numerical strings with different lengths, one 35 isolated letters as one sample of each isolated letter and the first 14 words of an Arabic word dataset – shown in Figure 1. The second page includes the rest of the candidate words as shown in Figure 2.

The forms have been filled by participants in two countries. The first one was filled in Montreal Canada, by 100 randomly selected Arabic writers from different genders, ages, educational levels and nationalities. The second form was filled in Saudi Arabia by 228 randomly selected participants. In the second form more words were added. Participants were asked to write the samples within the box boundaries using dark pen and try to make the images quite readable. For this reason; the databases have been divided into different series: 1 and 2, and 3. SERIES_1 contain the samples of the first 100 writers. SERIES_2 contains the samples from the last 228 writers. Finally, SERIES_3 contains the combination of samples from series 1 and 2. Different experiments could be conducted from different series. In total, there are 656 pages of filled forms.

4. Data Extraction and Pre-processing

After the forms were filled by the writers, they were scanned in the true color images of 300 dpi resolution. First, special filter was applied to remove all the red borders of the boxes that contained the target images. After removing all boxes from all of the forms, the true color forms were converted to greyscale forms. A special program has been developed to automate the data extracting process. In the design stage, four black boxes have been added to the corners of the forms. The program uses the coordinates of these boxes first to re-skew the form if it needs to, and to then locate the target areas of the forms. For each handwritten sample, the box's coordinates are located. After extracting all the handwritten samples, a special filter has been applied to remove the salt and pepper noise.

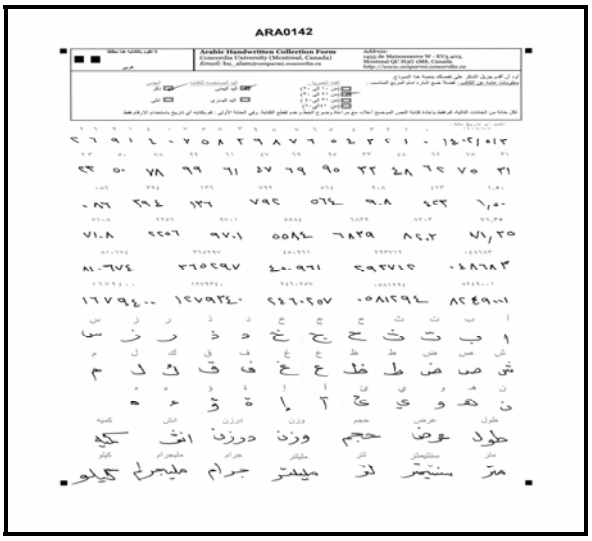


Figure1: Sample of filled form (Page 1)

Figure2: Sample of filled form (Page 2)

5. Database Overview

The general structure of the database is shown in Figure 3. The database is divided into three series: SERIES_1, SERIES_2 and SERIES_3. All the three sets share the same structure. SERIES_1 and SERIES_2 have completely different sets of images (no image appears twice in any set, so different experiments can be applied to each dataset) and SERIES_3 includes the union of both SERIES_1 and SERIES_2. Every series has the greyscale and the binary versions of the samples. Each series consists of five basic data sets: Dates, Isolated Letters, Numerical Strings, Words, and Special Symbols. The last part of the database is the FORMS folder which contains the contents of the original data-entry forms that have been used to collect the samples. Both the binary and the greyscale versions of the forms are included. In

the following section, a complete description and the statistics of each data set are presented.

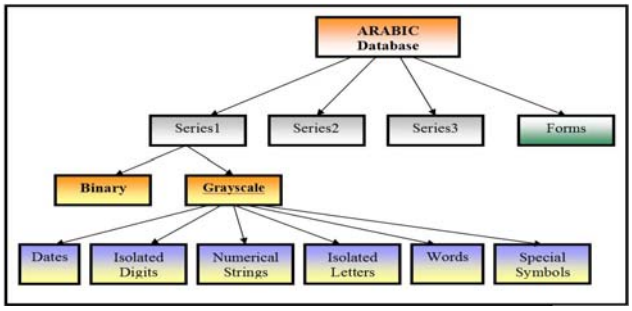


Figure 3: The general structure of our Arabic Database

5.1. Arabic Date Dataset

In the Arabic language, there is no one standard format for writing dates. The purpose of creating this dataset is to capture the different styles and formats for writing dates in Arabic. Therefore, one sample of a free-format date was included in the forms.

Some participants chose to write the date using the Arabic/Islamic (Hijri) calendar, while others chose to use the Gregorian calendar. All of the different samples have been accepted as long as they were written in Indian digits (Figure 5). Some of the participants who chose to write the date using the Arabic/Islamic calendar added the letter "Ha" "هـ" which represents the (Hijri) (هجري) calendar. The statistics for the Date dataset from series 3 is shown in Table 1.

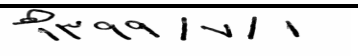
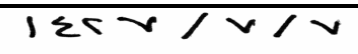
English Date	Arabic Date
H1399/7/1	
1427/7/7	

Figure 4: Different samples of Arabic dates. Different writers used different formats and different calendars. In the last row, the date ends with letter "Ha" "هـ" which represents that Arabic/Islamic calendar (Hijri) (هجري)

Number of Writers	Training Set	Verifying Set	Testing Set
284	170	57	57

Table 1: Statistics dataset for date

5.2. Indian Isolated-Digit Dataset

The data-entry form has two samples of each of the basic isolated Indian digits. In the numerical strings, each digit has been repeated 13 to 17 times at different positions; at the beginning, in the middle and at the end of the numerical strings. In order to gain an advantage of having a collection of numerical strings, special and advanced segmentation algorithm was used to extract the isolated digits from the numerical strings to create the isolated Indian digits database. A common problem in the Indian digits is the ambiguity in writing the digits 2 and 3 (Figure 5). Some statistics for this set from series 3 are shown in Table 2**.**

Handwritten Indian Digits				
Printed Indian Digits	٣	٣	٢	٢
Equivalent Arabic Digits	3	3	2	2
	(a)	(b)	(c)	(d)

Figure 5: Two different samples of the digits 3 and 2.From columns (b) and (c), we can see the ambiguity between 2 and 3.

Number of Writers	Training Set	Verifying Set	Testing Set
328	28000	9400	9400

Table 2: Statistics for Isolated Indian Digits Dataset

5.3. Indian Numerical Strings Dataset

During the collection process, each participant wrote 38 different numerical strings with different lengths (2, 3, 4, 6, and 7 digits per string). All the digits have been included in those numerical strings in all the different positions. Moreover, two decimal numbers were included. This dataset is divided into two sets: Integer set and real set. In the Integer set, different numerical strings with different lengths are included. The second set includes decimal numbers of the lengths 3, 4 and samples of the float points. Samples for the integer and real numeral strings are shown in Figures 6 and 7. Statistics for this database are shown in Table 3.

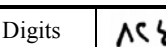
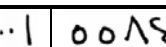
Handwritten Indian Digits		
Printed Indian Digits	٨٢٤٩٠٠١	٥٥٤٨
Equivalent Arabic Digits	**8249001**	**5584**

Figure 6: Samples of integer numeral strings

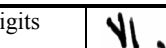
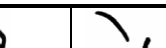
Handwritten Indian Digits		
Printed Indian Digits	٧١,٣٥	١,٥٠
Equivalent Arabic Digits	**71,35**	**1,50**

Figure 7: Samples of real numeral strings

Number of Writers	Training Set	Verifying Set	Testing Set
328	8033	2702	2704

Table 3: Statistics for the Indian Numerical Strings Dataset

5.4. Arabic Isolated Letters Dataset

The Arabic alphabet consists of 28 basic letters and there is no distinct upper and lower case letter form. Words in Arabic are written in a cursive manner, where most of the letters are directly connected to the letter that immediately follows. A few letters do not connect to the following letter, even in the middle of a word. Each individual letter can have up to four distinct forms, based on its position within a word or a group of letters - At the beginning, in the middle, at the end, or isolated.

The second important components that usually accompany some letters are called Harakat, which are vocalization diacritics that mark vowels and other sounds that can not be represented by Arabic letters. Hamzah is one the most important vocalization diacritics, which indicates a glottal stop and appears some times by itself or over other letters [3]. Moreover, Some Arabic letters can be written in different styles, therefore, collecting samples from those different styles is significantly important. As a result, the data-entry form includes one sample of the 34 Arabic isolated letters and one sample of Hamzah. Samples of those letters are shown in Figure 8, and statistics for this dataset are shown in Table 4.

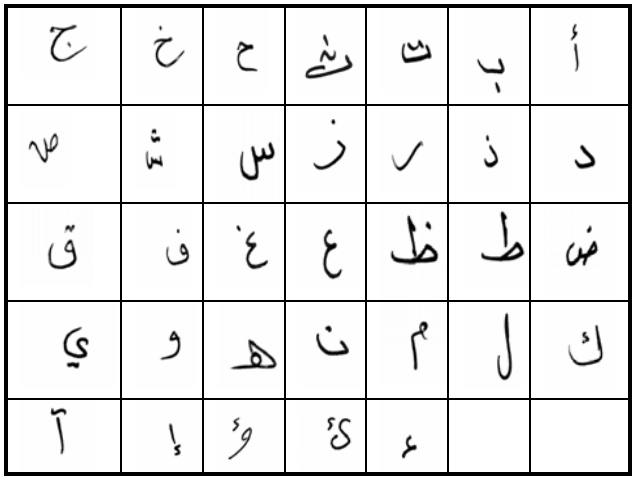


Figure 8: Sample of Handwritten Arabic Isolated Letters.

Number of Writers	Training Sets	Verifying Sets	Testing Set
328	12693	4367	4366

Table 4: Statistics for Arabic Isolated Letters Dataset

5.5. Arabic Words Dataset

A collection of 70 Arabic words has been selected to form the Arabic Word Dataset. The aim of creating this data set is to target new groups of words that have never been collected in any previous Arabic handwriting database. This data set could provide new and different challenges in the recognition of Arabic handwriting. The group includes: weights, measurements, and currencies. Samples of some of these Arabic words are shown in Figure 9. The currencies used in this database are the Saudi Arabian Riyal (SAR), and Hallalh, (1 Riyal = 100 Hallah). Statistics for this data set are shown in Table 5.

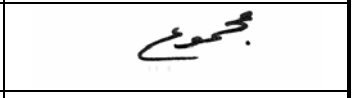
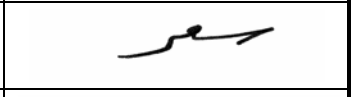
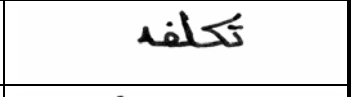
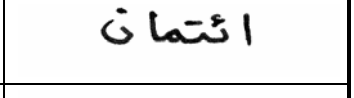
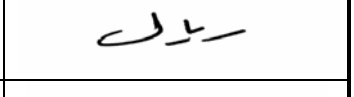
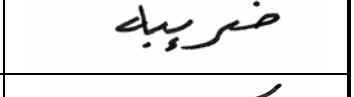
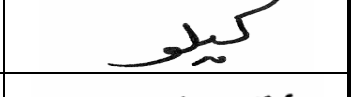
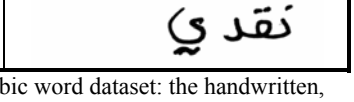
Sale	بيع	
Total	مجموع	
Price	سعر	
Cost	تكلفه	
Credit	ائتمان	
Riyal	ريال	
Gram	جرام	
Tax	ضريبه	
Kilo	كيلو	
Cash	**نقدي**	

Figure 9: Samples of Arabic word dataset: the handwritten, the printed, and the English words

Number of Writers	Training Set	Verifying	Testing Set
328	6790	2275	2310

Table 5: Statistics for words data set

5.6. Special Symbols Dataset

In addition to the Arabic words, the data-entry form includes general symbols that appear in any Arabic document. Although these symbols are non-language related, a small set is included in order to capture the style of writing for these symbols from Arabic native speakers. These samples are: comma (,), colon (:), at (@), slash (/), and no. (#). Statistics for this dataset are shown in Table 6.

Number of Writers	Training Set	Verifying Set	Testing Set
328	980	330	330

Table 6: Statistics for symbols dataset

6. Ground Truth Data

In the final structure of our Arabic database, each folder that contains handwritten samples is also provided with the ground truth data file for the samples. The ground truth data file includes the following information about each sample: image name, content, number of CCs (Connected Components), writer number, age, gender, and handwriting ordination. To the best of our knowledge, no Arabic handwriting database includes the writer gender, age, and hand-orientation in the ground truth data. An example of the ground truth data for the Date dataset is shown in Figure 10.

Image Name	ARA0274_P01_001.tif	ARA0109_P01_001.tif
Format	Yyyy/mm/dd	yyyy/mm/dd
Content	1426/6/6	1400/1/5
Writer No.	ARA0274	ARA0109
Gender	Female	Male
Hand Orientation	Right-hand	Left-hand
Age	41-60	31-40
Length(No. of CCs)	10	10
Image Type	DATE	DATE

Figure 11: Examples of the ground-truth data for Arabic date Dataset

7. Experiment Results

In order to show the usefulness of our database for research in Arabic handwritten recognition, we have conducted some experiments on segmentation and recognition of handwritten Arabic (Indian) numerals from our database, and the results are shown in this section. For our experiments in this section, first we considered isolated digits and then numeral strings from the first 100 writers in our database (Series #1). For the recognition of the isolated digits, similar features to the one used in [10] were extracted and they are shown in Figure 12. For the classifier also we applied a K-NN (K nearest neighbor classifier, with K=3). The overall results on isolated digits are shown in Table 7.

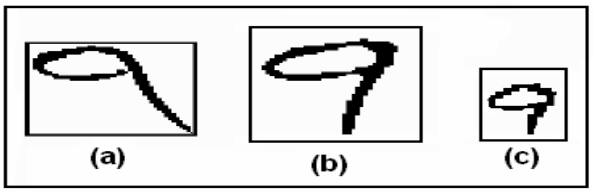


Figure 12: (a) Original image, (b) Smoothed, slant corrected, and normalized image (45 by 45 pixels), (c) Skeleton of part (b) is taken and its resolution is reduced in horizontal and vertical directions by down sampling (1/3). The resulting (15 by 15 pixels) image which is considered a (2D array) feature representation of the structure and style of writing of a digit.

Digits	0	1	2	3	4	5	6	7	8	9	Num. Cor.	Num. Err.	Num. Rej.
0	175	16	1	3	5	0	1	0	0	13	175	39	0
1	1	285	0	0	1	0	1	0	0	1	285	4	0
2	0	3	211	76	1	0	0	6	0	0	211	86	0
3	0	0	7	295	0	0	0	2	0	0	295	9	0
4	0	0	2	0	294	0	0	0	0	0	294	2	0
5	13	0	2	0	1	238	0	0	3	1	238	20	0
6	0	1	0	0	0	0	297	0	0	3	297	4	0
7	2	1	1	4	1	0	0	273	0	0	273	9	0
8	1	0	0	1	0	0	0	0	273	3	273	5	0
9	0	0	0	1	0	0	3	0	1	339	339	5	0

Table 7: Confusion Matrix on testing set (based on 100 writers in our database). The overall recognition rate on the testing set was 93.60%. Out of total 2863 samples of handwritten digits from 100 writers 2680 were correctly recognized and 183 were mis-recognized with zero rejections.

We also did some experiments for the segmentation and recognition of handwritten numeral strings. We chose randomly 754 handwritten numeral strings with different lengths (lengths: 2, 3, 4, 6, and 7 digits per string, the lengths were unknown to the segmentation algorithm) from the testing set of our database (based on 100 writes of Series #1) and we applied the segmentation algorithm developed in [11, 12] for segmenting those numeral strings. We also, applied the same isolated digit classifier explained above for the recognition of the segmented digits. Some results of these experiments are shown in Figure 13. In these experiments the accuracy at the digit level was 90.16% and at the level of the whole strings, the accuracy was 73.54%. This experiment shows sophisticated algorithm for segmentation of handwritten Arabic numeral strings should be developed in order to improve the accuracy.

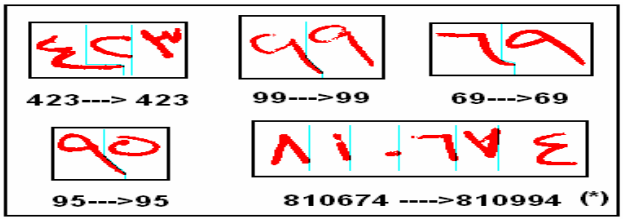


Figure 13: Results of combination segmentation and recognition of handwritten numeral strings, the sample marked with (*) shows a misrecognition case.

Conclusion and Future Works

With increasing interest in Arabic handwriting recognition [13][14], the need for a standard comprehensive Arabic handwriting benchmark database for researchers to compare their approaches is highly demanded. In this paper, we presented a new comprehensive database for Arabic handwriting recognition. It is the first database that covers six different important sets of the Arabic language: dates, isolated letters, isolated digits, numerical strings, a collection of 70 important words and a collection of special symbols. Developing this comprehensive database is an important step towards the expansion of Arabic handwriting recognition. The database will be made available in the future for research purposes from the Centre for Pattern Recognition and Machine Intelligence (CENPARMI), at Concordia University. In the future, we are looking to expand this database by adding more samples of Arabic words and adding samples of Arabic free texts and sentences.

Acknowledgement

Creating a multipurpose database requires a lot of effort and collaboration from many different people. We would like to appreciate all the people who helped us in achieving this goal. We are also thankful to all anonymous writers who spent their time and filled our data entry forms in Canada and in Saudi Arabia; also we appreciate Ms. Shira Katz for her editorial assistance. Furthermore, we sincerely appreciate the NSERC of Canada, FQRNT of Quebec, and CENPARMI, Concordia University for their valuable support of our research and we hope it will get popularity in the research community.

References

[1] A. El.Sagheer, N. Tsuruta, and R.-I. Taniguchi , *"Arabic Lip-reading System: A Combination of Hypercolumn Neural Network Model with Hidden Markov Model"*, Artificial Intelligence and Soft Computing ASC, 9.1 - 9.3, Marbella, Spain, 2004.

[2] F. Barbara, Ethnologue: Languages of the World, *14th ed: SIL International*, 2000.

[3] L.M. Lorigo, V. Govindaraju, "Offline Arabic handwriting recognition: a survey", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, pp. 712- 724, 2006.

[4] U. Marti, H. Bunke, "A full English sentence database for off-line handwriting recognition", *Proc. of the 5th Int. Conf. on Document Analysis and Recognition*, pp. 705-708, Bangalore, 1999.

[5] J.J. Hull, "A Database for Handwritten Text Recognition Research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 550-554 , 1994.

[6] G. Dimauro, S. Impedovo, R. Modugno, and G. Pirlo., "A new database for research on bank-check processing", *Proc. of the 8th Int. Workshop on Frontiers in Handwriting Recognition*, pp 524-528, Canada, 2002.

[7] M. Pechwitz, S. Maddouri, V. Märgner, N. Ellouze ,and H. Amiri, "IFN/ENIT-Database of Handwritten Arabic Words" , *The 7th Colloque International Francophone surl'Ecrit et le Document*, Tunis, 2002.

[8] S. Alma'adeed, D. Elliman, and C.A. Higgins, " A Database for Arabic Handwritten Text Recognition Research", *In Proc. Eighth International Workshop of Frontiers in Handwriting Recognition*, pp485-589, Canada, 2002.

[9] Y. Al-Ohali, M. Cheriet, and C. Suen, "Database for Recognition of Handwritten Arabic Cheques", *Proc. of the 7'Th Int. Workshop on Frontiers in Handwriting Recognition*, pp. 601-606, The Netherlands, 2000.

[10] J. Sadri, C. Y. Suen, and T. D. Bui., "A New Clustering Method for Improving Plasticity and Stability in Handwritten Character Recognition Systems", *Proc. of Int. Conference on Pattern Recognition*, vol.2,pp. 1130–1133, Hong Kong, 2006.

[11] J. Sadri, C. Y. Suen, and T. D. Bui., "Automatic Segmentation of Unconstrained Handwritten Numeral Strings", *In Proc. of Int. Workshop on Frontiers in Handwriting Recognition*, pp. 317–322, Japan, 2004.

[12] J. Sadri, C. Y. Suen, and T. D. Bui., "A Genetic Framework Using Contextual Knowledge for Segmentation and Recognition of Handwritten Numeral Strings", *Pattern Recognition*, Vlo.40, pp. 898–919, 2007.

[13] R. Al-Hajj, C. Mokbel, and L. Likforman-Sulem, "Combination of HMM-Based Classifiers for the Recognition of Arabic Handwritten Words", *Proc. of the 9Th Int. Conference on Document Analysis and Recognition*, Vol. 2, pp. 959-963, Brazil, 2007.

[14] S. Touj, N. Ben Amara, and H. Amiri, "A hybrid approach for off-line Arabic handwriting recognition based on a Planar Hidden Markov modeling", *Proc. of the 9Th Int. Conference on Document Analysis and Recognition*, Vol. 2, pp. 964-968 , Brazil, 2007.