A New Multipurpose Comprehensive Database for Handwritten Dari Recognition

Muhammad Ismail Shah *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)-848-2830.
Emails: {mu_shah, j_sadri, suen, nicola}@cse.concordia.ca

Abstract

In this paper, we present the creation of the first comprehensive database for research and development on handwritten recognition of Dari language. This new handwritten database consists of many aspects of Dari scripts such as: handwritten isolated characters, isolated digits, numeral strings of various lengths, many words/terms, dates, and some special symbols. For each handwritten image in this database, very useful ground truth information is provided to facilitate successful recognition experiments on the images. The data has been archived into two different formats - Gray level and Binary. The contents of the database are frequently used in several kinds of documents such as scientific and business documents. The overall structure of the database has been designed in such a way to make it convenient for conducting recognition experiments on the handwritten Dari scripts.

Key Words: Optical Character Recognition (OCR), Handwritten Recognition, Dari Handwritten Database, Dari Handwritten Recognition, Farsi and Arabic Hand-written Recognition.

1. Introduction

This paper describes the creation of the first comprehensive and standard handwritten database of Dari language. Dari, also known as Persian, is one of the Indo-Iranian languages, a subfamily of the Indo-European languages [1]. Dari and Pashto are the two official languages of Afghanistan but Dari is considered to be the lingua franca for the languages of Afghanistan [4]. Approximately five million people in Afghanistan and a total of two and half million people in Iran, Pakistan, and neighboring regions, speak Dari [4]. In addition, in other parts of the world such as North America, Australia and many European countries, there are also hundreds of thousands of Dari speakers living as refuges/immigrants [1].

Dari and Farsi languages use almost the same type of modified Arabic script, however, in Dari the stress accent is less prominent as compared to Farsi [1]. Hence, the main difference between Dari and Farsi is in the grammar. In addition, Dari also has some loan words from Arabic and Farsi languages. Dari is also a cursive language like Arabic, Farsi and Pashto and is written from right to left [1], except the numerals which are, like Farsi, written from left to right. Furthermore, Dari not only uses the Farsi numerals but also shares the same 32 letters with Farsi language. Due to this fact this new database may prove to be a very useful source for Farsi handwritten recognition research.

These days many people and organizations are interested in the Optical Character Recognition of handwritten scripts of the Indo-Iranian languages i.e. Arabic and Farsi and some useful databases have been created for the last several years. For example, [2] is a handwritten collection of data about Farsi isolated digits, English isolated digits, isolated alphabets, Farsi dates and Farsi legal amounts which is useful for handwritten check recognition research. ‘CENPARMI Arabic cheques’ [3] is another database which contains Arabic hand printed bank cheques and is useful for cheque recognition researches.

For Dari language, however, no such a useful and standard handwritten database has been reported so far by the research communities. Hence, we made some efforts at the Center of Pattern Recognition and Machine Intelligence (CENPARMI), Concordia University, Canada to create a useful handwritten database for Dari language. The result of these efforts is the creation of very standard and multi-aspect handwritten database for Dari language. We hope that this database will not only promote the study and research in handwritten recognition of these challenging cursive scripts, but also will be very useful for the companies who are developing Optical Character Recognition technologies for cursive scripts, i.e. South-Eastern Asian Scripts.

This database contains various aspects of Dari scripts such as characters/alphabets, isolated digits, numeral strings, dates, some special symbols, and many useful Dari words and terms related to science, business, finance documents as well as cursive legal amounts. All the handwritten data has been pre-processed and distributed into three groups i.e. Training, Testing and Validation Sets, accompanied by proper and useful ground truth information which makes our database very useful for handwritten recognition research.

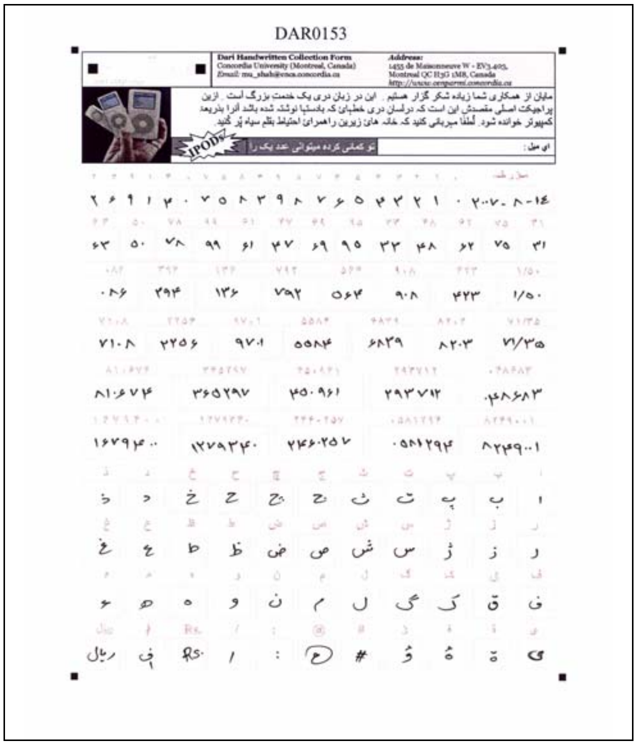
The rest of the paper has been organized in such a way that Section 2 describes the process of data collection, Section 3 briefly explains the extraction and archiving of handwritten data, Section 4 gives a general overview of the structure and contents of the database, Section 5 shows a sample of the ground truth information for each image, Section 6 gives the overall statistics of the handwritten data while Section 7 presents results of some segmentation and recognition experiments and in Section 8 we have given some conclusive remarks and future works.

2. Data Collection

2.1. Data Collection Forms

The tool used for data collection purpose is a Data Entry Form which consists of two pages. The first page as shown in Figure 1 was used for collecting handwritten Dari dates, isolated digits, numeral strings, isolated letters and some special symbols. The second page, as shown in Figure 2, was used for collecting Dari handwritten words relating to various subjects. The second page also contains a section, in the top which gathers information about the gender and hand-orientation of the writer.

Both pages of the data entry form were filled by the same writer are identified by using an ID which uniquely specifies the writer. This ID consists of 7 characters of which the first three are always 'DAR' which stands for Dari while the remaining four digits specifies the writer number who filled the form.


DAR0153

Dari Handwritten Collection Form
Concordia University (Montreal, Canada)
Email: mu_shah@encs.concordia.ca

Address:
1455 de Maisonneuve W - EV3.403,
Montreal QC H3G 1M8, Canada
http://www.cenparmi.concordia.ca

ما از همکاری شما زیاده شکر گزار هستیم . این در زبان دری یک خدمت بزرگ است . ازین
پراجیکت اصلی مقصدش این است که درلسان دری خطهای که بادستها نوشته شده باشد آنرا بذریعه
کمپیوتر خوانده شود. لطفاً میربانی کنید که خانه های زیرین را همراه احتیاط بقلم سیاه پُر کنید

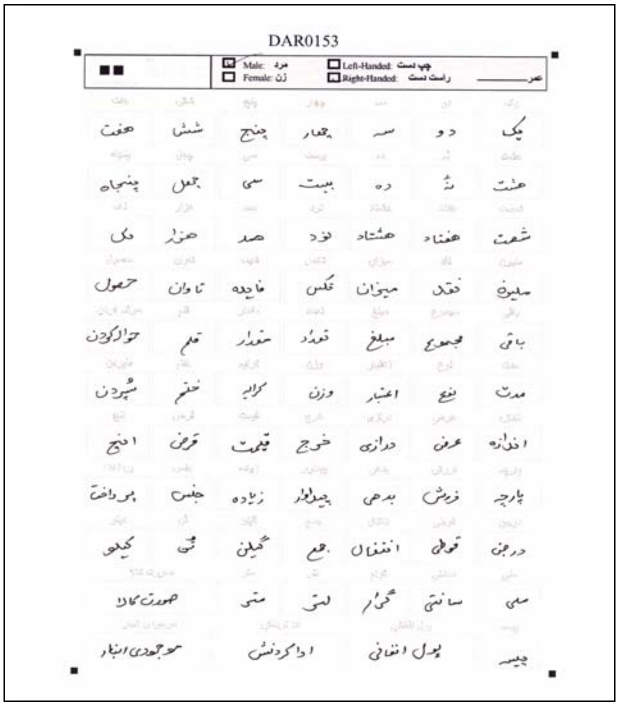
تو کمکی کرده میتوانی عدد یک را

ای میل :

۲۰۰۷-۰۸-۱۴ ۰ ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۳ ۸ ۵ ۷ ۰ ۴ ۱ ۹ ۶ ۲

۳۱ ۷۵ ۶۲ ۴۸ ۳۳ ۹۵ ۶۹ ۴۷ ۶۱ ۹۹ ۷۸ ۵۰ ۶۳

Figure 1: A Sample of the filled Data Entry Form, Page 1.


DAR0153

☒ Male: مرد ☐ Left-Handed: چپ دست
☐ Female: زن ☒ Right-Handed: راست دست عمر ______

یک دو سه چهار پنج شش هفت

هشت نه ده بیست سی چهل پنجاه

Figure 2: A Sample of the filled Data Entry Form, Page 2.

2.2. Data Gathering

After designing our forms, the process of data collection has been mainly conducted in Swat district of North West Frontier Province (NWFP) of Pakistan. This district has a large number of Dari speakers, coming from Afghanistan, and are living as immigrants or students. The handwritings of 200 native Dari writers/speakers belonging to various educational and socio-economic backgrounds were collected. Overall, based on the gender and hand-orientation, we have four groups of Dari writers: right-handed males, right-handed females, left-handed males and left-handed females.

3. Extraction and Storage of Data

3.1. Scanning and Pre-processing of Entry Forms

All the filled forms have been scanned in true color format (24 pits per pixel) then converted into gray level format (8 bits per pixel) and have been archived in TIFF (Tagged Image File Format) images with 300 dpi resolution. After scanning, all the forms were pre-processed to remove the type written labels, colored lines and noise that is usually caused by the scanner (such as noisy strips on the margins of images etc).

3.2. Extraction and Pre-processing of Data Fields

The required data fields on the scanned forms were extracted automatically through some developed computer programs. After extraction of all the fields, they were pre-processed in order to remove all the remaining salt and pepper noises using a 3 by 3 Median filter [6]. Furthermore, the bounding boxes of all the fields have been adjusted. All the extracted images have been archived in two different formats: gray scale and binary formats in TIFF file format in 300 resolutions. We selected TIFF file format for archiving the images because it is a standard format and is widely accepted these days [3].

4. Overview of Database Contents

4.1. Dates in Dari language

The date format in Dari language is similar to that of Farsi and Pashto languages which is usually written as yyyy/mm/dd (year/month/day). This is considered as the standard format of date in Dari and Farsi, and we used this format to label dates in our data collection form. An example of handwritten Dari date is shown in Figure 3.

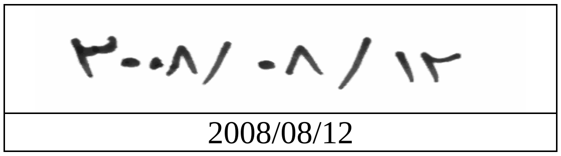


Figure 3: An example of handwritten Dari date.

4.2. Special Symbols

Our database also contains 7 special symbols of which 4 are related to the computer science that include the “at” symbol, number sign, slash and colon. The handwritten samples of these symbols are shown in Figure 4.

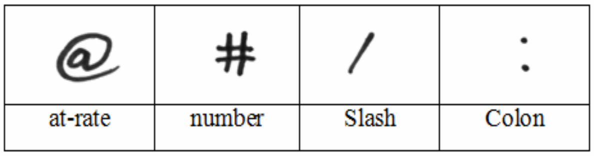


Figure 4: Computer related symbols

Furthermore, one of these symbols i.e. slash is also used to represent the decimal point in Dari language. The other three symbols are related to currency signs of official currencies of Afghanistan, Pakistan and Saudi Arabia.

4.3. Dari Words

In Dari, like Farsi and Arabic scripts, words are written in cursive forms. Our database contains a set of 73 Dari words which are frequently used in various kinds of documents such as scientific and business documents. In addition, it also contains many words used for writing legal amounts in bank checks. An example of some of these handwritten words is shown in Figure 5. These handwritten cursive words can be used for various kinds of segmentation and recognition experiments.

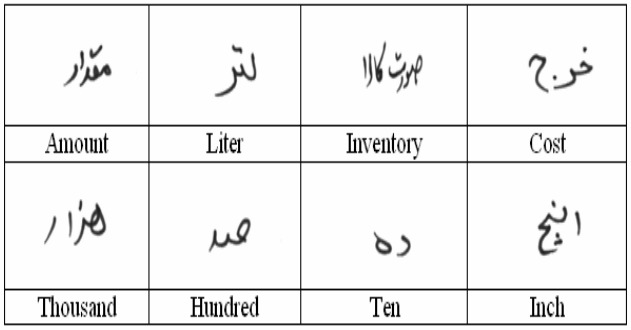


Figure 5: The various types of Dari words

4. 4. Dari Alphabets

The Dari language is the Afghan dialect of Farsi, so it uses the same 32 alphabets/letters of Farsi language. In addition to these 32 letters we also considered five other letters, three of which are the modified forms of letter ه i.e. ة, ۀ, and هـ. The letter هـ is the shape of ه at the initial position within a word. The other two added letters include ؤ (waw hamza) and Arabic letter ء (Hamza Alef). Every writer wrote these 37 letters only once in the Data Entry Forms. Figure 6 shows hand written samples of these Dari letters, taken from a filled form.

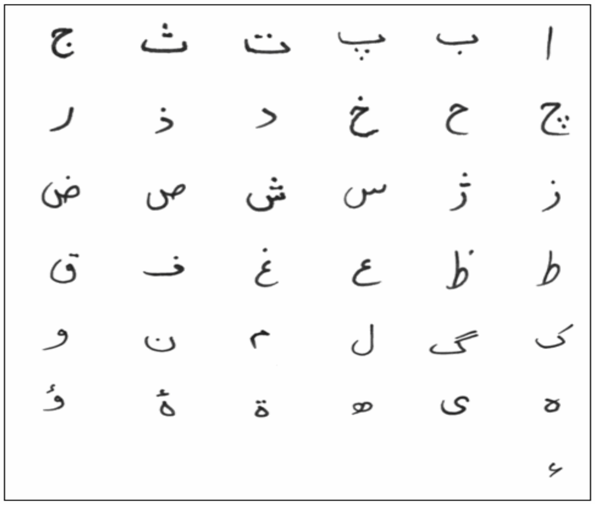


Figure 6: Handwritten Dari Isolated Letters

4.5. Dari Isolated Digits

Similar to the alphabets, the Dari digits are also very similar to Farsi digits. Figure 7 shows the handwritten Dari Digits along with the corresponding labels.

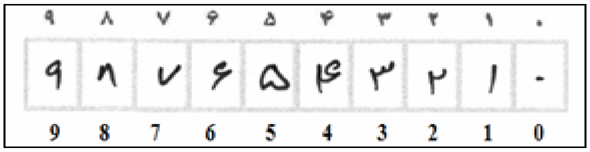


Figure 7: Dari Isolate Digits

In our database every writer has written each digit 14 times on average (2 times in isolated form, and 12 times from handwritten numeral strings which have been segmented by the segmentation algorithm described in [8][9]). We also found several variations in the styles of writing of handwritten digits in our database. These styles of writing are normally used in other scripts such as Arabic and Urdu. For example, for digits 0, 2, 4, 5, 6 and 7 we have seen variations as shown in Figure 8.

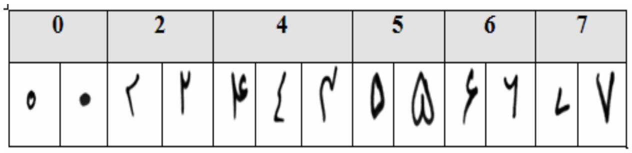


Figure 8: Variations in Handwritten Isolated Digits

4.6. Dari Numeral Strings

In our data entry forms as shown in Figure 1, there are 38 different fields containing handwritten numeral strings. These numeral strings have various lengths from 2 to 7 digits. Out of these 38 numeral strings, 36 are integer strings (they don't contain decimal point), and two are real strings (they contain decimal point). Out of the 36 integer strings, thirteen strings are of length 2, seven strings had length 3, six strings are of length 4, five strings are of length 6 and five strings are of length 7. Some of the handwritten samples of these numeral strings are shown in Figure 9.

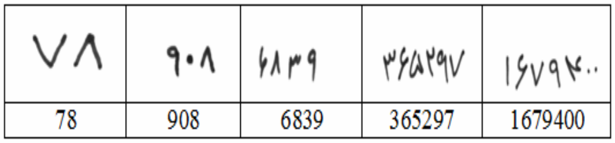


Figure 9: Handwritten Samples of Integer Strings

The two numeral strings which contain decimal point are shown in Figure 10. As shown in Figure 10, the decimal point in Dari language is represented by a small slash.

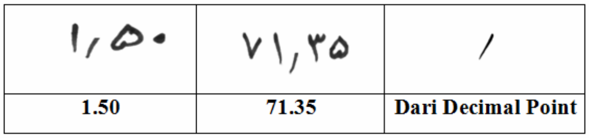


Figure 10: Dari real strings and isolated decimal point

5. The Ground Truth Information

Accurate and well structured ground truth information is considered to be and essential part of a handwritten database which makes it more useful and convenient for conducting experiments. For each type of handwritten data (six types, as discussed in Section 4) in our database, we have provided a set of ground truth information which has been stored in their corresponding folder for each type. Table 1 shows an example of our novel ground truth information for an image of a numeral string in our database. As shown in Table 1, for each image some important information such as image name, writer ID, writer's gender, hand-orientation, data type, content, and length have been provided. These ground truth information have been provided in the form of text files, which are compatible to almost all platforms, e.g. Windows, Linux, etc… operating systems.

Table 1: The Sample Ground Truth Information

Image	
Image Name	**DAR0015_P01_055.tif**
Writer's ID	**DAR0015**
Writer's Gender	**Male**
Hand-Orientation	**Right-Handed**
Data Type	**Integer String**
Label (Content)	**8249001**
# of CCs (Length)	7

6. Training, Testing, and Validation Sets and Overall Statistics of the Database

In order to make this database ready for recognition experiments, the data has been divided into three disjoint sets: Training, Testing, and Validation Set. This distribution was randomly done in such a way that 60% of all the images were assigned to Training set, while 20% to Testing and Validation sets each. Table 2 summarizes the total number of handwritten samples for each type of data and its distribution among Training, Testing and Validation sets.

Table 2: Overall statistics

Type of Data			Number of Classes	Total Number	Training Set	Testing Set	Validation Set
Dates			1	200	120	40	40
Isolated Characters			37	7400	4440	1480	1480
Isolated Digits (0-9)			10	28000	16800	5600	5600
Words			73	14600	8760	2920	2920
Special Symbols			7	1400	840	280	280
Numeral Strings	Integer Strings	Length 2	13	2600	1560	520	520
		Length 3	7	1400	840	280	280
		Length 4	6	1200	720	240	240
		Length 6	5	1000	600	200	200
		Length 7	5	1000	600	200	200
	Real Strings	Length 4	1	200	120	40	40
		Length 5	1	200	120	40	40

7. Experimental Results

In order to show the usefulness of our database, we have conducted some experiments on segmentation and recognition of handwritten numeral strings. For our experiments in this section, first we trained an isolated digit classifier on the training set of our Dari isolated digits (1680 digits per 10 classes). For the classifier we used a K-NN (K nearest neighbor classifier, with K=3), and for the features we used a similar method to [7]. Figure 11 shows the pre-processing and feature extraction for our classifier.

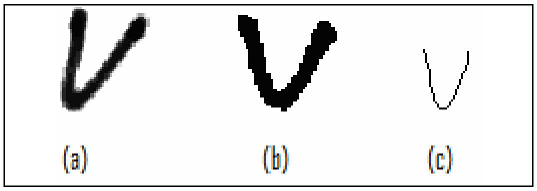


Figure 11: (a) Original image, (b) Normalized, slant corrected, and smoothed image (45 by 45 pixels), (c) Skeleton of part (b) is taken and its resolution is reduced in horizontal and vertical directions by down sampling (1/3). The resulting (15 by 15 pixels) image which is considered a (2D array) feature representation of the structure and style of writing of a digit.

Using our classifier and the segmentation algorithm as described in [8] [9], we conducted some experiments for segmentation and recognition of handwritten numeral strings. We took in total 480 numeral strings of different lengths (4 and 7 digits per strings) from the testing set of our database. Figure 12 shows some results of segmentation and recognition of these Dari numeral strings. Out of these 480 numeral strings 91.45% were segmented correctly, and at the digit level, the classifier could recognize 84.67% of the digits in the segmented numeral strings. This experiment shows that a sophisticated algorithm for segmentation of handwritten Dari numeral strings should be developed in order to improve the accuracy of segmentation and recognition.

8. Conclusion and Future Work

In this paper, we have described the creation of the first Comprehensive handwritten database for research and development on handwritten recognition of Dari language. Our database includes many aspects of Dari script such as: handwritten isolated characters, isolated digits, numeral strings (different lengths), words, dates, and handwritten special symbols. In our database, for each image the ground truth information such as writer information (unique ID, gender, and hand-orientation), data type, true content (true label), and the number of connected components have been provided which make this database to be very valuable and convenient for any research experiments on Dari script. All the images are available in two different formats: gray level and binary. The overall structure of this database has been designed in such a way that it is very convenient for training, testing, and validation of different algorithms for segmentation and recognition of many aspects of Dari language. In future, we are going to expand the number of words in our database and add free hand written texts of Dari to this database and do more experiments on segmentation and recognition of numeral strings and cursive words. This database will be made available in the future, for research purposes. We hope that this database will become popular and will be useful for researchers who are interested in handwritten recognition of cursive scripts such as Dari and other related languages such as Farsi and Arabic.

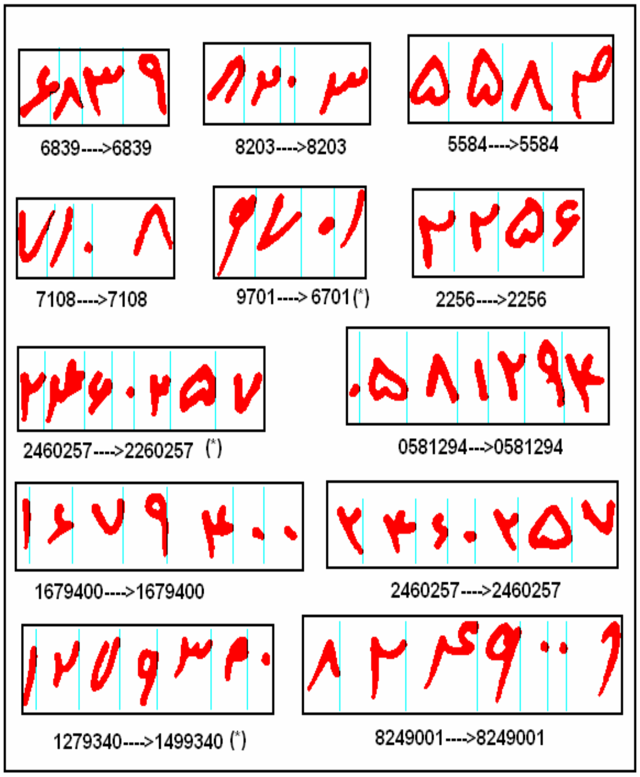


Figure 12: Some results of our segmentation and recognition algorithm are shown here. The samples shown with (*) are misrecognized samples.

Acknowledgement

Creating a database needs a lot of efforts and collaboration of many people; we would like to pay our gratitude to all the people who helped us in achieving this goal. Especially, we are greatly thankful to Mr. Fazal Hadi Sarhadi who helped us to collect this handwritten data in Pakistan. We are also thankful to all those writers who filled our data entry forms.

References

[1] en.wikipedia.org/wiki/Dari_language

[2] F. Solimanpour, J. Sadri, and C. Y. Suen, "Standard Databases for Recognition of Handwritten Digits, Numerical Strings, Legal Amounts, Letters and Dates in Farsi

Language", *Proceedings of the 10th Int'l Workshop on Frontiers in Handwriting Recognition*, La Baule, France, Oct. 2006, pp. 3-7.

[3] Y. Al-Ohali, M. Cheriet, and C.Y. Suen, “Databases for recognition of handwritten Arabic cheques,” *Proceedings of the Seventh Int. Workshop on Frontiers in Handwritten Recognition*, Amsterdam, The Netherlands, Sep 2000, pp. 601-606.

[4] en.wikipedia.org/wiki/Tagged_Image_File_Format

[5] www.lmp.ucla.edu/Profile.aspx?LangID=191&me

[6] W.K. Pratt, *Digital Image Processing*, Wiley, New York, 1978

[7] 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, Hong Kong, 2006, pp. 1130–1133.

[8] J. Sadri, C. Y. Suen, and T. D. Bui, “Automatic Segmentation of unconstrained Handwritten Numeral Strings”, *In Proceedings of International Workshop on Frontiers in Handwriting Recognition*, Tokyo, Japan, October 2004, pp. 317–322.

[9] J. Sadri, C. Y. Suen, T. D. Bui, "A Genetic Framework Using Contextual Knowledge for Segmentation and Recognition of Handwritten Numeral Strings", *Pattern Recognition*, v.40 n.3, March 2007, pp. 898-919.