ARABASE: A Relational Database for Arabic OCR Systems

Najoua Ben Amara [1], Omar Mazhoud [1], Noura Bouzrara [2], and Noureddine Ellouze [2]
[1] National School of Engineer of Monastir, Tunisia
[2] National School of Engineer of Tunis, Tunisia

Abstract: *In this paper we present a database for the research of Arabic off-line and on-line handwriting optical recognition as well as for machine printed text optical recognition. Digital images of documents, text phrases, words/sub-words, isolated characters, digits, signatures, soon are and included in ARABASE. Data corresponds to a variety of lexes (cities names, literal amounts, isolated characters, digits, free texts, etc.). The database organization offers interesting commodities to be explored via an Arabic writing recognition system. A useful tool enables the user, via a graphical interface to experiment different classical tasks of image processing.*

Keywords: *Databases, Arabic writing recognition, on-line and off-line handwriting, printed documents, multi-fonts, multi-writers.*

Received July 2, 2004; accepted September 17,2004

1. Introduction

Through Arabic Optical Character Recognition (AOCR) dates back to the 70's, it remains unsolved. Among this subject, related most crucial problems, we should mention the lack of test tools, particularly databases, vocabularies, validation dictionaries and statistics about Arabic (bi-grams, tri-grams and n-grams). These tools are of a great necessity to the improvement of OCR systems performances. They allow the validation of the results of recognition systems. They also permit to evaluate on common resources, the new approaches and algorithms that are proposed in different data processing environments and to perform benchmarks and thereby measure progress and general tendencies [19]. In the field of Arabic recognition, a few privately owned databases are frequently used they are often collected in a laboratory environment which makes AOCR products difficult - even impossible- to compare. Absence of Arabic characters calligraphies normalization, from a typographical point of view, makes comparing more problematic [3, 5]. Consequently, the actual published recognition performances have to be considered as valid only concerning the test vocabulary, often limited (a unique font, restricted number of writers, good quality documents…). In this context, performances have announced in AOCR related works and already published a really objective significance?.

In this work, we propose a database for the research of Arabic off-line and on-line handwriting optical recognition as well as for machine printed text optical recognition [6]. We will pay interest particularly to the description of this database and its different functionalities.

This paper is organized as follows. In the next section, we briefly recall the main characteristics of Arabic writing so as to well understand the upheld choice at the level of ARABASE architecture. We present in section 3 a state of the databases developed for the research in AOCR. In section 4, we describe the ARABASE architecture, along with the different data, which it contains. The different functionalities offered are briefly discussed in section 5. Conclusions are addressed in section 6.

2. Characteristics of Arabic Writing

Arabic has a different writing when compared to other writings in different ways. One of the differentiations comes from its semi-cursivness aspect in both its handwritten and printed forms, hence the existence of the sub-word. A sub-word is then a chain of connected characters. Thus the Arabic word is consequently composed of one or more sub-words. We also keep in mind the strong dependence of the shape of a character on its context, the complexity of the letterform, the variability of inter-characters linking, both horizontal and vertical.

Two up to four close characters may combine vertically to give birth to very particular drawings (called vertical combinations) whose morphology is mostly different from characters, which constitute them. Printed Arabic is known for its richness in fonts and styles. More than 450 styles and fonts-all different exist. From a font to another, the morphological

characteristics of the Arabic character change considerably [10].

Handwritten recognition is a challenging problem especially in the case of Arabic due to its calligraphic nature [4].

Figure 1 gives examples of printed and handwritten Arabic writing.

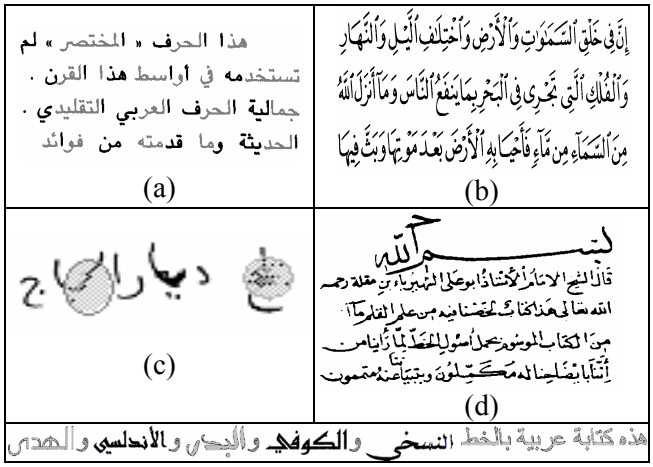


Figure1. Illustrations of a few characteristics of Arabic writing: (a) the different sub-words of a printed text, (b) voweled writing, (c) examples of overlapping characters, (d) an example of handwritten writing, (e) multi-font aspect, [4].

3. Database in AOCR

Databases containing a significant representative number and selection of samples constitute a potential resource. They are necessary tools to the different experiments of OCR systems as well as to the advanced research in the field of writing optical recognition. In fact, these bases allow us to evaluate, upon common resources, the performances of different approaches and the recorded error analysis permits to deduce future developments in the subject matter.

Ideally these bases should be issued from real documents. They are as important as their big size. In fact, the number of samples used during the training stage affect perceptibly the performances of an OCR system. To train an OCR system with a reduced number of samples limits its later capacity of recognition. Nevertheless, the OCR oriented databases are hard to conceive at the individual researchers level, taking into account the underlying fastidious tasks and the costs involved, particularly at the level of data collecting and their digitization. Moreover and very often, arises the problem of confidentiality of certain data, i. e., the writer's name and signature, hence the importance of developing data for researcher's use in the field.

Different list databases for different OCR systems exist such as Latin and Japanese, which contribute considerably in the advance of their OCR research. Some of these bases are public, a few of which are free, but some others are commercialised. We give as examples the following databases: NIST [18], CEDAR [12], UNIPEN [11], IRONOFF [19] and GRUHD [13]. Unfortunately, such bases are nearly absent in the Arabic case. Consequently, most AOCR researchers have to gather data individually and hence, they test their systems on "private" images, which constitute a big inertia to the development of efficient methodologies in the field.

Three databases have recently had birth; they are offered to the use of researchers. They are: The Tunisian-German IFN/ENIT database, the CEMPARMI database (Concordia University Montreal), and finally that of the University of British Columbia, Canada. More recently a database of Arabic handwriting was proposed by Al-Ma'adeed *et. al.* [1]. The database contains the most popular words in Arabic writing.

We should mention that the former bases are not real databases as the name means, but rather merely a collection of organized labelled samples of images. In the following, we briefly describe the content of the previous first three data sets.

3.1. IFN/ENIT Database

The Tunisian Signal and Processing Systems Laboratory in collaboration with the German Institute of Telecommunication has provided AOCR researchers with an image dataset, which corresponds to the names of Tunisian town/villages. Those names are multi-writers handwritten. 411 writers -who belong to different age and cultural range- have participated in the collection of the data and at present, 26, 459 names and zip codes of city names have been collected [15]. The different images are scanned with 300 dpi and converted to binary images. Each handwritten town name comes with image and ground truth information (postcode, Arabic word in ISO8859-6 code set, number of words, number of character, writer identifier…). In addition, the IFN/ENIT database contains 10 frequently used vertical ligatures. The IFN/ENIT is publicly available for the purpose of research on the Internet on the following site: www.ifnenit.com/. It is broadcast free to any interested researcher.

3.2. CENPARMI Database

In collaboration with Al Rajhi Banking and Investment Corporation (one of the largest banking corporations in Saudi Arabia), the CENPARMI laboratory has developed real-life databases for Arabic legal amounts and courtesy amounts (written in Indian digit) [2]. About 7,000 real world grey-level cheque images were collected and previously deprived of all personal information including names, account numbers and signatures). The dataset is sub-devised into various sub-bases. The first sub-base numbers 2, 499 legal amounts; the second one includes 2, 499 courtesy amounts written in Indian digits. The third sub-base is

composed of 29, 498 sub-words within the domain of legal amount. The last sub-base contains 15, 175 Indian digits. Each of these sub-bases is divided into training (66% to 75%) and testing sets. The different databases are in TIFF format. In [2] we do find a well-detailed description of this basis. This dataset is commercialised.

3.3. British Columbia Database

In the Computer Science and Electric Engineering Department in the British Columbia University, Canada, a dataset of phrases, handwritten Arabic words, digits and signatures has been developed [14]. 500 randomly selected students have participated in the data collection which took place in Al-Isra' University in Amman, Jordan. Each student has been requested to sign 5 times and copy a pre-selected list of words and digits, as well as a sentence. The basis words were scrupulously chosen in a way to insure the presence of all the Arabic alphabet letters in their various letterforms.

The basis contains 37, 000 Arabic words, 10,000 Indian and Arabic digits, 2, 500 signatures and 500 free form Arabic sentences; all saved in grayscale and black and white .bmp file formats and ready for use.

However, the datasets, which contain city names or literal amounts or even addresses, are insufficient. They can be aimed only to a specific country. In fact, city names are strictly proper to a specific country and cannot be explored to another country. Also, the literal amounts can be applied only to a particular currency in a particular country. Consequently, the application that comes from such databases can't be exploited, in real applications, outside that country. In addition to that, digits used in the Arabic world may be either Indian or Arabic. Hence, AOCR researchers are bound by the restricted vocabulary of such bases. In addition, according to their contents, all these datasets can be used only in the research of Arabic off-line handwriting.

Owing to all these reasons, we have suggested the development of a database, open enough which may be exploited in local or other applications by collecting data (printed as well as on-line and off-line handwritten) of a different nature (texts, words, isolated characters, digits, signature…), leaving the field open to different types of tests and applications.

4. Description of ARABASE

ARABASE is a rich database, including images associated to data of different nature. A friendly user-interface under windows 98/2000 and XP allows the user to have access to the different recordings of the basis. It allows the visualization of the various entities contained in the basis and information to which they are associated. The visualization tool permits user to apply a range of images processing operations. Similarly, different statistics related to the basis vocabulary may be achieved.

In the following, we firstly consider the data collection phase; secondly, we describe the basis architecture, and the data organization.

4.1. Collecting Data

It is a critical phase. The major problem which arises at this level is the gathering of data issued from real documents in order to target concrete applications such as postal sorting or the recognition of literal amounts. However, this task is problematic owing to the difficulty met in acquiring real data; for instance, post offices and banks are constrained by the laws, which ensure their client's confidentiality when dealing with this information.

ARABASE contains a wide data issued from machine printed and handwritten documents. The collection of off line-handwritten data has been operated particularly in distant academic environments. Right now more than 400 writers have participated in the collection of the data. Most writers were Tunisian. We use different A4-forms with predefined boxes. The content of these forms was inspired from that used in the case of the UNIPEN data set [11]. The writer was asked to fill in the empty space boxes, copying the text just above (digits, city names, amounts…), a space box was left for the signature (Figure 2). The ground truth of the text to be written is provided in machine print. Each filled form is then verified by a human operator to ensure that it has been correctly filled. In addition to this, many city names extracted from the IFN/ENIT dataset were similarly introduced in our database.

Regarding the printed data, the different documents introduced up to now correspond to images of texts selected from a range of daily newspapers and books published on Internet by the Tunisian National Library [20]. We have also introduced texts selected from the Tunisian Official Journal that we have retyped in different fonts and different sizes, and printed using different machines (laser and ink jet printers).

The various images (except those already published electronically), printed as well off-line handwritten, were scanned as grayscale, black and white images using two scanners with a resolution ranging from 200 to 600 dpi.

4.2. ARABASE Organization

4.2.1. Data Organization

For each recognition area (printed recognition and on-line and off-line handwriting recognition) ARABASE is subdivided into several sub-bases each representing a specific entity of data (city names, literal amounts, digits, isolated characters…). To facilitate the results sharing among researchers as well as performance

comparisons, each sub-base is found in a separate directory containing the testing images samples (~1/3) and that of training (~2/3).

For each given entity sample composed by a chain of characters, we find its image as well as its image previously segmented (manually) into characters (Figure 3). This choice has been taken into account in order to facilitate and assist some procedures of segmentation while pointing out the different approximately limits among the characters.

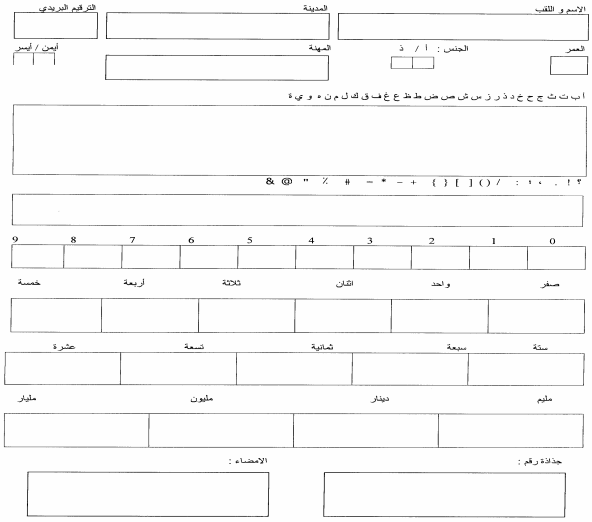


Figure 2. Example of a form to fill in, case of off-line handwriting.

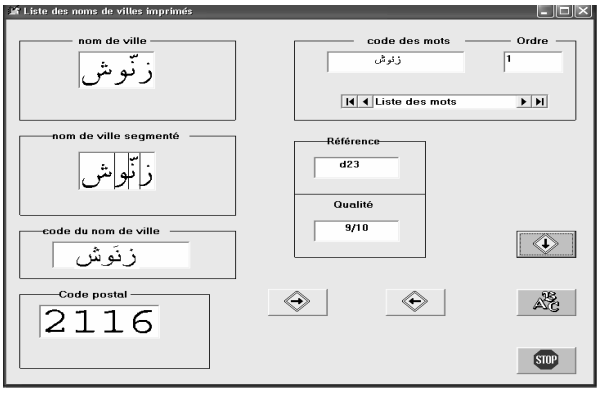


Figure 3. Screen page corresponding to the case of a printed document associated to a city name.

4.2.2. Description of ARABASE

A great effort has been made at the level of the conceptual model development of our database. The task is not trivial owing to the diversity of the considered pieces of information and their inter-relations. In fact, ARABASE contains data associated to multiple pieces of information: Tunisian city names, zip codes, literal amounts, drawings of overlapping characters, digits signatures, symbols and free texts (some of which are voweled) extracted from documents of different origins (books, newspapers). Figure 4 gives an extract of the physical model of ARABASE.

As Figure 5 shows, ARABASE is structured around the concept of "document" which represents the data highest level. The different levels of the basis correspond, conforming to the: Descending segmentation principle, text entities (city names, literal amounts…), words, sub-words, and characters, which compose them. The document corresponds to the original document (for instance the filled forms in the case of off-line handwritten), from which the various entities are extracted.

The document is described mainly with four attributes: its reference code, type, quality and context. The reference code or the ID represents the main key, which indicates the document name. The different documents are sorted according to their reference codes. The type specifies the writing mode: Printed, handwritten on-line or off line. The quality corresponds to a value ranging from 0 to 10, attributed to the document according to its degradation degree. The context specifies the description of the data contained in the document: Literal amounts, city names, digits, symbols, and vertical linkages.

With the exception of the vertical linkages, each of the entities is segmented into sub-entities (words, sub-words, characters). The segmentation procedure is presently supervised.

The entity and the sub-entities are described by the following attributes: Their coding "UNICODE", their image and their segmented image (see Figure 3).

In the case where the main entity corresponds to vertical linkages, i. e., to overlapping characters, we also find the corresponding image of the succession of the untied characters (Figure 6). This allows facilitating at the same time the procedure of segmentation and that of recognition of an AOCR system.

Besides, and depending on the writing mode, whether printed or handwritten on-line or off line, more pieces of information are available.

4.2.2.1 Case of the Printed

Bearing in mind the big number of fonts in the case of Arabic printing [9], we have thought beforehand to specify, among the attributes of printing, the font name and the point size used in the document. More over, we have kept the printing characteristics (resolution and the type printing: Offset processes, inkjet, laser or dot matrix) and those of digitization (scanner's resolution and speed). These characteristics are important, they determine how well experimental results on that data set can be generalized to another application [11].

4.2.2.2 Case of the off-line Handwritten

Other than the digitization characteristics, information associated to the writer is taken into account. It deals mainly with the name, first name, age, sex, job and signature of the writer (Figure 7).

4.2.2.3 Case of on-line Handwritten

The characteristics of the acquisition tool (e. g., the discretized pen trajectory information, the pressure of the pen, the characteristics of the digital tablet: Speed, sampling frequency, the spreadsheet dimensions…) are retained as well as the information related to the writer, such as his signature.

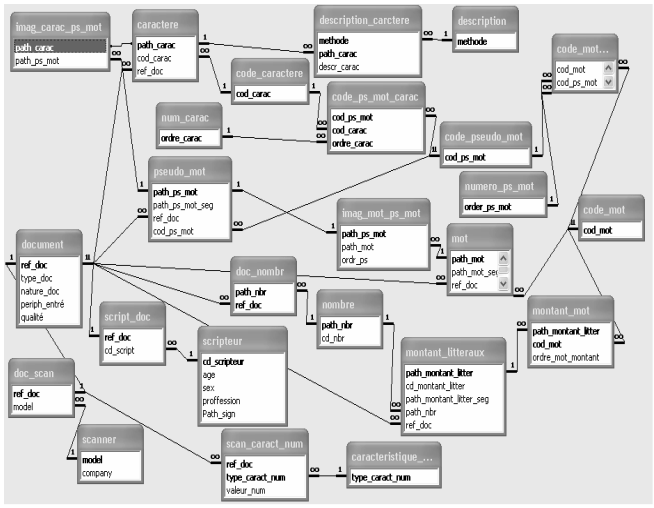


Figure 4. An extract of the physical model of ARABASE.

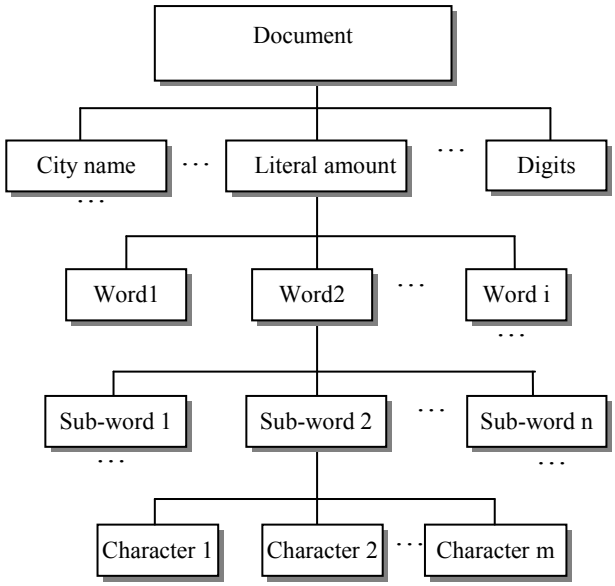


Figure 5. The general structure of a document.

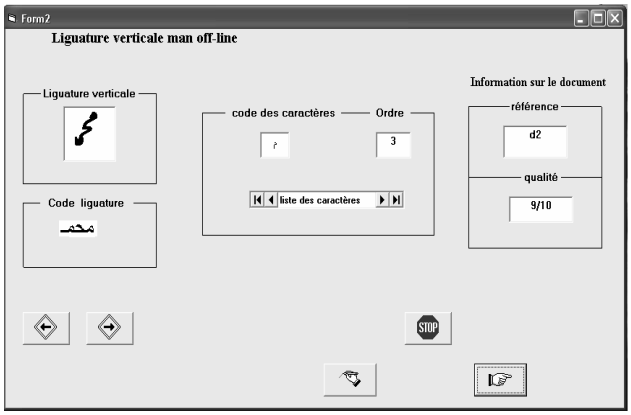


Figure 6. A screen page in the case of overlapping characters.

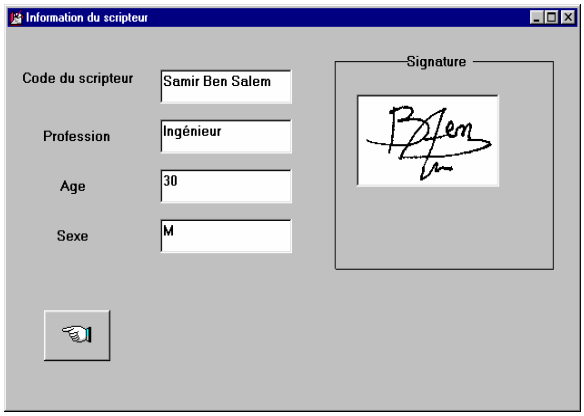


Figure 7. A screen image giving the writer's characteristics (off-line case).

5. Functionalities of ARABASE

A database for the research writing recognition is essentially meant to feed the training phase of the system and its test phase. Apart from these tasks, we have thought of the basis so as to offer other functionalities, which aim at assisting the recognition procedure and well exploit to the best the data in which they are introduced. Different requests SQL have then been defined. They were effortlessly implemented, thanks to the basis architecture. The results of these different requests are saved in tables, which are directly accessible through an AOCR system. Among the most important requests we mention (Figure 8):

- The consultation by entity (or sub-entity).
- The OCR assistance at the level of segmentation as well as at the level of recognition, offering respectively the approximate bounds among characters and a limited index of the most plausible candidates.
- The statistics relative to the different bi-grams, tri-grams or even n-grams (at level of the characters, sub-words and words).
- Pre-processing of the basis images: The graphical ARABASE interface is equipped with an Image-Processing (IP) module allowing to carry out a set of operations of IP on the different images of the basis.

A degradation model is also provided for users, it is intended for generating degraded synthetic images of the different images entities included in the database.

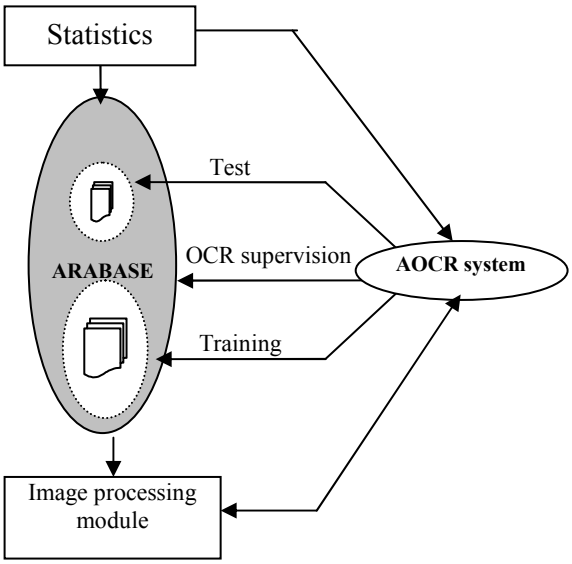


Figure 8. The main functionalities of the basis and interfacing with an AOCR system.

6. Conclusions

In this paper, we have addressed the critical problem of lack of databases in AOCR field. We have discussed three datasets already developed for Arabic recognition and shown their limits for a wide use application.

We have described ARABASE, a database for the research of Arabic off-line and on-line handwriting optical recognition as well as for machine printed text optical recognition. This basis offers images that correspond to multiple pieces of information leaving the field open to diversified applications. ARABASE addresses various research considerations. Apart from the classical functions of training and of testing, the information structuring as a database in the informatics' meaning offers several opportunities such as targeted vocabulary consultation, an AOCR assistance system during the segmentation phase and/or recognition, images pre-processing operations, statistics…

The diversity of the data nature (city names, literal amounts, signatures, vertical linkages, free texts…) leaves the field open to training and testing of vast panoply of applications (postal sorting, recognition of literal amounts, forms, signatures identification writers…).

ARABASE is already three years old [6, 7, 8], and collecting data is still in progress in order to enlarge the size of ARABASE.

On-line data is being collected using a sensitive pad, our object is to offer a database that can be shared with all AOCR researchers.

Besides, we are experimenting the different functionalities of ARABASE on two AOCR systems, which are being developed by our team. The first AOCR system is conceived for the recognition of printed Arabic writing in consideration of an approach based upon the Generalized Hough Transform [16]. The second system deals with handwritten Arabic writing based on Planar Hidden Markov Models [17]. The various tests show the good structuring of the basis and the efficiency of its various functionalities for use in a context of Arabic recognition.

In order to make the use of ARABASE more convivial, we foresee to develop a segmenter so that a new document will be automatically segmented and introduced to the database.

References

[1] Al-Ma'adeed S., Elliman D., and Higgins C., "A Database for Arabic Handwritten Text Recognition Research," *International Arab Journal of Information Technology*, vol. 1, no. 1, pp. 118-123, January 2004.

[2] Al-Ohali Y., Cheriet M., and Suen C. Y., "Databases for Recognition of Handwritten Arabic Cheques," *Pattern Recognitio*n, vol. 36, no. 1, pp. 111-121, 2003.

[3] Ben Amara N. and Bouslama F., "Classification of Arabic Script Using Multiple Sources of Information: State of the Art and Perspectives," *International Journal on Document Analysis and Recognition*," vol. 5, no. 4, pp. 213-223, July 2003.

[4] Ben Amara N. and Ellouze N., "Overview and Advances in Arabic Optical Character Recognition," *Asian Journal on Information Technology*, Grace Publications Network, vol. 3, no. 4, pp. 217-240, 2004.

[5] Ben Amara N., "Sur la Problématique et les Orientations en Reconnaissance de l'écriture Arabe," *in Proceedings of Colloque International Francophone sur l'Ecrit et le Document (CIFED'02)*, Hammamet, Tunisia, pp. 1-10, 2002.

[6] Ben Amara N., Mazhoud O., Bouzrara N., and Chatti R., "ARABASE: Une Base de données Pour la Recherche en Reconnaissance de l'écriture Arabe," *in Proceedings of IEEE, Sciences of Electronic, Technologies of Information and Telecommunications (SETIT'2004)*, Tunisia, March 2004.

[7] Boulehmi H., "Développement d'une Base de Données Test Pour un Système AOCR," *Engineering Project*, National School of Engineer of Monastir, Tunisia, June 2002.

[8] Chatti R. and Bouzrara N., "Base de Données Pour un Système AOCR," *Master Project*, Faculty of Sciences of Monastir, Tunisia, 2003.

[9] Gazzah S. and Ben Amara N., "Une Approche Neuronale d'identification des Une Approche Neuronale d'identification des Fontes Arabes," *in Proceedings of IEEE, Sciences of Electronic Technologies of Information and*

Telecommunications (SETIT'2004), Tunisia, March 2004,.

[10] Gazzah S. and Ben Amara N., "Utilisation des Ondelettes en Caractérisation des Fontes Arabes," *in Proceedings of International Conference on Image and Signal Processing (ICISP'2003)*, Marroco, 2003.

[11] Guyon I., Haralik R. M., Hull J. J., and Phillips I. T., "Data Sets for OCR and Document Image Understanding Research," in Buke H. and Wang P. S. P. (Eds), *Handbook on Optical Character Recognition and Document Image Analysis*, World Scientific Publishing Company, 1996.

[12] Hull J. J., "A Database for Handwritten Text Recognition Research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550-554, May 1994.

[13] Kavallieratou E., Liolios N., Koustsogeorgos E., Fakotakis N., and Kokkinakis G., "GRUHD: Agreek Database of Unconstrained Handwriting," *in Proceedings of IAPR, ICDAR'01 6th Conference on Document Analysis and Recognition*, Seattle, Washington, USA, pp. 561-565, 2001.

[14] Kharma N., Ahmed M., and Ward R., "A New Comprehensive Database of Handwritten Arabic Words, Numbers and Signatures Used for OCR Testing," *in Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering*, Edmon, Canada, 1999.

[15] Pechwitz M., Maddouri S. S., Margner V., Ellouze N., and Amiri H., "IFN/ENIT-Database of Handwritten Arabic Words" *in Proceedings of the Colloque International Francophone sur l'Ecrit et le Document (CIFED'2002)*, Hammamet, Tunisia, pp. 129-136, 2002.

[16] Touj S., Ben Amara N., and Amiri H., "Generalized Hough Transform for Arabic Optical Character Recognition," *in Proceedings of IAPR of the 7th International Conference on Document Analysis and Recognition (ICDAR'2003)*, Edinburg, Scotland, pp. 1242-1246, August 2003.

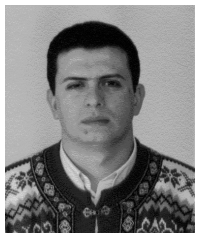
[17] Touj S. M., Ben Amara N., and Amiri H., "Segmentation Stage of a PHMM-Based Model for Off-Line Recognition of Arabic Handwritten City Names," *in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC 2002)*, October 2002.

[18] Wilkinson R., Geist J., Janet S., Grother P., Burges C., Creecy R., Hammond B., Hull J., Larsen N., Volgl T., and Wilson C., "Fist Census Optical Character Recognition Systems Conference," *Technical Report NISTIR 4912*, The U.S Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD, 1992.

[19] WWW.bibliotheque.nat.tn, 2004.

[20] Viard-Gaudin C., Lallican P. M., Knerr S., and BinterP. , "The IRSTER on/off (IRONNFF) Dual Handwriting Database," *in Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR'99*), Bangalore, India, pp. 455-458, 1999.

Najoua Ben Amara received her BSc, MSc, and PhD degrees in electrical engineering, signal processing and system analysis, from the National School of Engineer of Tunis, Tunisia, in 1985, 1986 and 1999, respectively. From 1985 to 1989, she was a researcher at the Research Institute on Informatics and Telecommunication IRSIT, Tunis, Tunisia. In September 1989, she joined the Electrical Engineering Department of the National School of Engineer of Monastir, Tunisia, as an assistant professor. In 2004, she obtained the PhD supervision degree and currently she is an associate professor at the same institution. Her research interests include mainly optical character recognition applied to Arabic documents, image processing, compression, the use of stochastic models, wavelets, and hybrid approaches in the above applications.



Omar Mazhoud received his BSc in computer science in 1990 from the Faculty of Economic Sciences and Management of Sfax-Tunisia. In 1991, he received his MSc in computer science from the Paul Sabatier University of Toulouse, France. In 1999, he joind the National School of Engineer of Monastir, Tunisia as an assistant professor. He is now preparing a PhD on the annotations and the computer-aided reading activities themes.

Noura Bouzrara received her MSc in computer science in 2003 from the Faculty of Sciences of Monastir, Tunisia. Currently, she is a PhD student in the research group of Signal Processing Research Laboratory at the National School of Engineer of Tunis, Tunisia. She is working on the topic of databases for Arabic OCR systems.

Noureddine Ellouze received a PhD degree in 1977 from l'Institut National Polytechnic at Paul Sabatier University (Toulouse-France), and Electronic Engineer Diploma from ENSEEIHT in 1968 from the same University. In 1978 He joined the Department of Electrical Engineering at the National School of Engineer of Tunis (ENIT–Tunisia), as assistant professor in statistic, electronic, signal processing and computer architecture. In 1990, he became a professor in signal processing, digital signal processing and stochastic process. He has also served as director of Electrical Department at ENIT from 1978 to 1983, general manager and president of the Research Institute on Informatics and Telecommunication IRSIT from 1987-1990, and president of the institute in 1990-1994. He is now the director of Signal Processing Research Laboratory LSTS at ENIT, and he is also in charge of control and signal processing master degree at ENIT. Dr Ellouze is IEEE fellow since 1987, he directed multiple masters and thesis and published over 200 scientific papers both in journals and proceedings. He is the chief editor of the scientific journal Annales Maghrébines de l'Ingénieur". His research interests include neural networks, fuzzy classification, pattern recognition, signal processing and image processing applied in biomedical, multimedia, and man machine communication.