AMHCD: A Database for Amazigh Handwritten Character Recognition Research

Youssef Es Saady, Ali Rachidi, Mostafa El Yassa, Driss Mammass
IRF-SIC Laboratory,
University Ibn Zohr, Agadir, Morocco

ABSTRACT
In this paper, we describe the first version of a database that contains handwritten Amazigh characters (AMHCD). As present the database consists of 25,740 isolated and labeled Amazigh handwritten characters produced by 60 writers. This database has been developed at the IRF-SIC Laboratory of the university Ibn Zohr, Agadir, Morocco. It is designed for training and testing recognition systems for handwritten Amazigh characters. This database is available for researches and academic uses.

General Terms
Pattern Recognition, Document Image Analysis

Keywords
Amazigh, Handwriting, Database, Amazigh OCR, Recognition system.

1. INTRODUCTION
Optical Character Recognition (OCR) has been a popular research area for many years because of its various application potentials, such as bank cheque processing, postal automation, documents analysis, etc. Several scientific researches have been carried out for character recognition of English language, Chinese languages, Arabic language, handwritten numerals, etc. and various approaches have been proposed by the researchers for automatic recognition of these characters. Thus, to standardize and compare research results, many databases in the handwritten recognition domain have been gathered and used in various languages and applications. There are databases in Latin [1-3], Chinese [4], Indian [5], Korean [6], Arabic [7-9] and Farsi [10] for offline handwritten recognition applications. However, only a few studies are attested on handwritten characters of Amazigh scripts. Recently, some efforts have been reported in literature for Amazigh characters based on the Hough transformation [11], the statistical and geometrical approaches [12], artificial neural network [13], [14], Hidden Markov Models [15], syntactical method based on the finite-state machines [16] and dynamic programming [17]. Besides, as far as we know, there has no standard database for Amazigh characters, which allows objective comparisons between different systems. All published works in [11-17] have been tested on local databases, which contain a restricted number of Amazigh characters. With the exception of the database of the Amazigh printed graphic developed in [13], which contains about 20000 printed characters. We used this database to test our approach presented in [14]. As part of the research on recognition of Amazigh characters, we developed an Amazigh Handwritten Character Database (AMHCD). This database will be intended to serve other researchers in the field and standardize the research on Amazigh character recognition. It contains more than 25000 images of handwritten Amazigh characters that comprise 33 classes.

The rest of the paper has been organized as: The next section presents an overview on the Amazigh language. 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 data storage and labeling are presented. Finally, we present conclusion and future work.

2. THE AMAZIGH LANGUAGE
The Amazigh (Berber) language is spoken in Morocco, Algeria, Tunisia, Libya, and Siwa (an Egyptian Oasis); it is also spoken by many other communities in parts of Niger and Mali. It is used by tens of millions of people in North Africa mainly for oral communication and has been introduced in mass media and in the educational system in collaboration with several ministries in Morocco. In linguistic terms, the language is characterized by the proliferation of dialects due to historical, geographical and sociolinguistic factors. In Morocco, the term Berber (Amazigh) encompasses the three main Moroccan variants: Tarifite, Tamazighte and Tachelhite. More than 40% of the country's populations speak Berber. So, all the Moroccans are concerned with this alphabet. The establishment of The "Royal Institute of the Amazigh Culture" (IRCAM) carried out a major action to standardize the Amazigh language. In the same tread, and since 2003, the Amazigh language has been integrated in the Moroccan Educational System. It is taught in the classes of the primary education of the various Moroccan schools, in prospect for a gradual generalization at the school levels and extension to new schools [18].

The Tifinagh is the writing system of the Amazigh language. An older version of Tifinagh was more widely used by speakers of North Africa. It is attested from the 3rd century BC to the 3rd century AD. The Tifinagh has undergone many changes and variations from its origin to the present day. The Libyan is the earliest varieties of Tifinagh. The Sahara Tifinagh is additionally called Libyan-Berber or old Touareg. For Tifinagh Touareg, differences exist in this alphabet of the value of symbols used by each population dialect. The Neo-tifinagh refers to the writing systems that were developed to represent the Maghreb Berber (Amazigh) dialects. The Neo-Tifinagh script was developed and computerized in the 20th century mainly by Moroccan and Algerian researchers, some of whom were based in Europe. The most important aspects of each of these variations are presented in [16]. The old Tifinagh script is found engraved in stones and tombs in some historical sites in northern Algeria, in Morocco, in Tunisia, and in Tuareg areas in the African Sahara. The figure 1 below presents a picture of a old Tifinagh script found in site of rock carvings near from Intedeni Essouk in Mali [19].

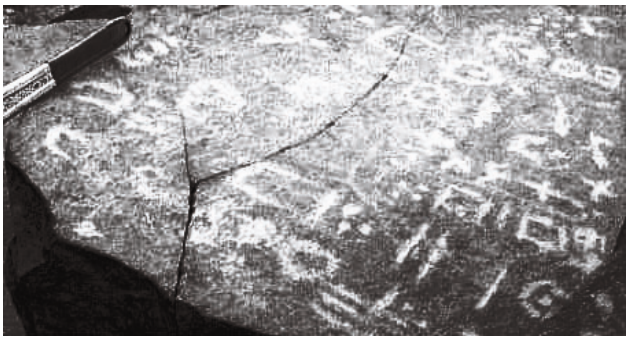

Figure 1: Old tifinagh script, site of rock carvings near from Intédeni Essouk Mali [19]

The Amazigh alphabets, called "Tifinagh-IRCAM", adopted by the Royal Institute of the Amazigh Culture, was officially recognized like belonging to the basic multilingual planned by the International Organization of Standardization (ISO)[20]. The Table 1 represents the repertoire of Tifinagh which is recognized and used in Morocco with their correspondents in Latin characters. The number of the alphabetical phonetic entities is 33, but Unicode codes only 31 letters plus a modifier letter to form the two phonetic units: ⵇ^{w} (g ^{w}) and ⴽ^{w} (k ^{w}).

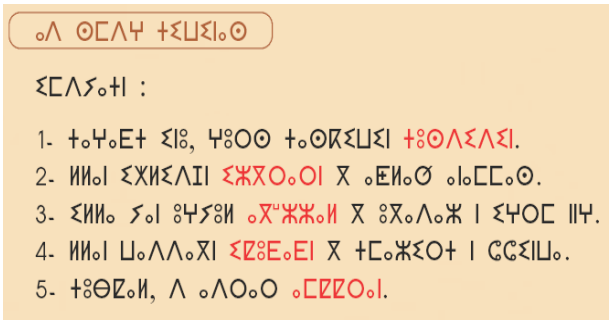

Figure 2: An example of an Amazigh text from a schoolbook

In contrast to Latin and Arab, the Amazigh alphabet is never cursive which facilitates the operation of segmentation. The Amazigh script is written from left to right; it uses conventional punctuation marks accepted in Latin alphabet. Capital letters, nonetheless, do not occur neither at the beginning of sentences nor at the initial of proper names. So there is no concept of upper and lowercase characters in Amazigh language. Regarding the figures, it uses the Arabic Western numerals. The majority of graphic models of the characters are composed by segments. Moreover, all segments are vertical, horizontal, or diagonal. Figure 2 above show some of these characteristics in a few Amazigh texts from a schoolbook.

Table 1: Tifinagh-IRCAM characters with their correspondents in Latin characters

Tifinaghe Pronunciation	Tifinaghe	Latin Correspondence
ya	ⴰ	a
yab	ⴱ	b
yag	ⴳ	g
yag^{w}	ⴳⵯ	g^{w}
yad	ⴷ	d
yaḍ	ⴹ	ḍ
yey	ⴻ	y
yaf	ⴼ	f
yak	ⴽ	k
yak^{w}	ⴽⵯ	k^{w}
yah	ⵀ	h
yaḥ	ⵃ	ḥ
yaɛ	ⵄ	ɛ
yax	ⵅ	x
yaq	ⵇ	q
yai	ⵉ	I
yaj	ⵊ	j
yal	ⵍ	l
yam	ⵎ	m
yan	ⵏ	n
yu	ⵓ	u
yar	ⵔ	r
yaṛ	ⵕ	ṛ
yaγ	ⵖ	γ
yas	ⵙ	s
yaṣ	ⵚ	ṣ
yac	ⵛ	c
yat	ⵜ	t
yaṭ	ⵟ	ṭ
yaw	ⵡ	w
yay	ⵢ	y
yaz	ⵣ	z
yaẓ	ⵥ	ẓ

3. COLLECTION OF DATA

Our database of isolated Amazigh handwritten characters was collected from 60 peoples. These writers were selected from various age, gender, and educational background groups. The samples were gathered by asking the informants to write on a form of 13 examples for each Amazigh character. We collected

420 documents, an example of a filled form used for collection of data is shown in figure 3 below.

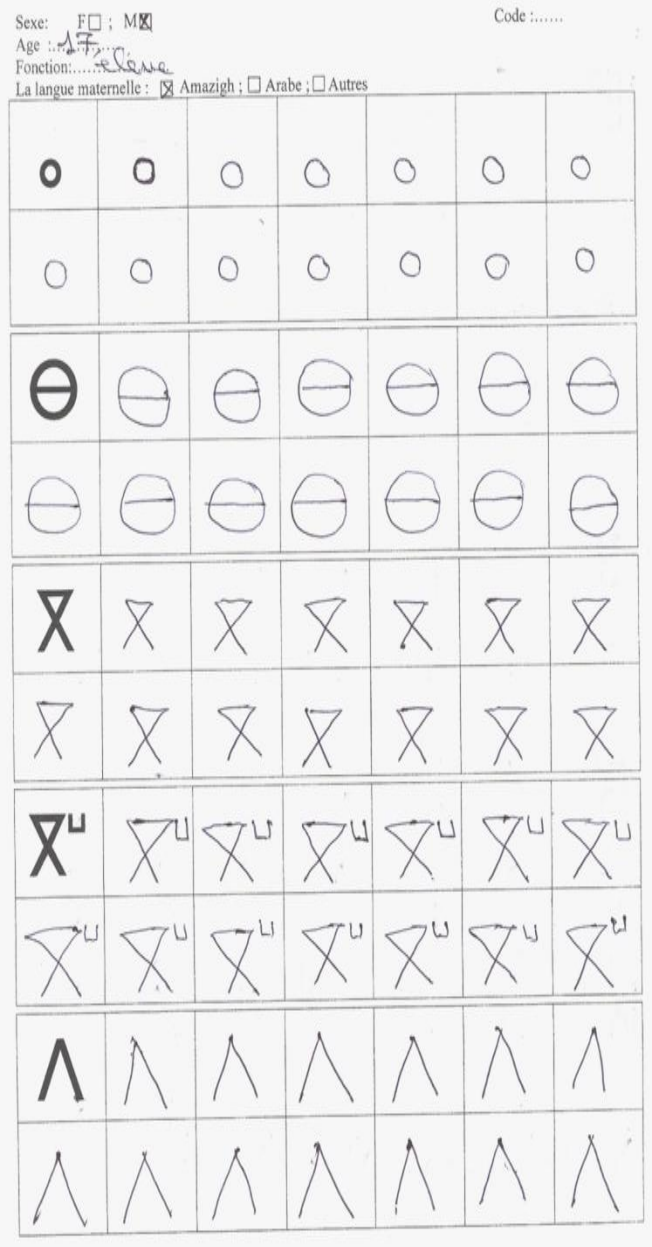

Figure 3: An example of a filled form

3.1 Form Design

We chose an input page layout that makes segmentation relatively easy, to avoid the complex problem of document segmentation in characters. We designed a form consisted from seven pages. The first page includes: some information about the writer in the header block and 5 isolated Amazigh characters as 13 samples of each isolated character as shown in Figure 3, and the other six pages contain 13 empty samples of each isolated character to be filled.

3.2 Forms Scanning

The collected documents are scanned using the HP-scan jet 5550c at 2400-dpi, which is usually a low noise and good quality image. The digitized images are stored as color images in the JPG format. The figure 3 above provides samples of a form scanned.

4. DATA EXTRACTION AND PRE-PROCESSING

After the digitizing of the forms collected, we have developed an automatic system to process and segment them into isolated characters. Indeed, a page slope correction was performed automatically using the Hough transform to estimate the skew angle and correct the skew of the scanned images [21]. Next, we developed an advanced horizontal/vertical projection method for automatic extraction of original isolated characters from the scanned forms. This method locates the characters of rectangular boxes in such images of the form by identifying a pair of parallel lines which are cut by several cells. We checked each sample manually for some segmentation errors; as a result all the characters in the database were well segmented. Then a normalization of the characters image is applied to eliminate unwanted areas using the projections techniques as shown in figure 4.

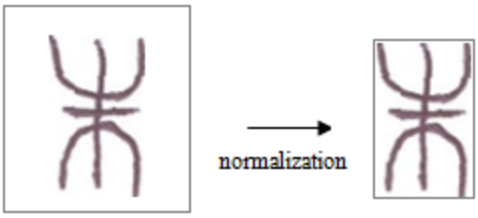

Figure 4: Normalization

One of the characteristics of the Amazigh script is that all characters are written as uppercase characters apart from the character ya (a) which is smaller than the others. Besides, the character ya (a) is very similar to the character yar (r), which is distinguishes by its size: the character ya (a) is a small circle while yar (r) is a big one. Sometimes, there is confusion between the images of these two characters. As a result, the previous normalization applied to the character ya (a) may generate problems. To overcome this problem, we added an empty space at the top of the character ya (a) after its normalization as shown in Figure 5. The size of this zone added is equal to the size of the normalized character ya (a).

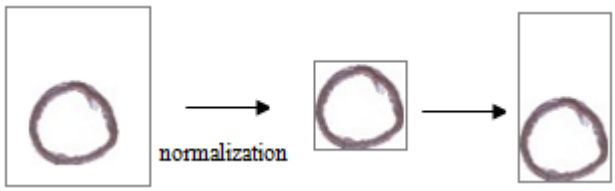

Figure 5: The normalization specified at the character 'ya'

5. DATA STORAGE AND LABELING

The extracted image is saved as the PNG format and named according to the following criteria; the database contains a list of

directories and each one of these directories represents a letter of the alphabet. Figure 6 below displays a few folders of the database. These directories are named by the Tifinagh pronunciation of the contained character according to the table 1. Therefore, any newly acquired image file should be added to one of these directories. The name of the image file within its corresponding directory is formatted as follows: characterPronunciation_writerNb_sampleNb.FileExtension; characterPronunciation is the Tifinagh pronunciation of the abbreviated character; writerNb is the number of writer; sampleNb is the sample number written by the writer and FileExtension is the extension of the file being stored.

Figure 6: Few folders of the database AMHCD

A few samples of isolated Amazigh characters from the present database are shown in table 2.

6. CONCLUSION & FUTURE WORK

A first version of the AMHCD database has been presented in this paper. This database contains more than 25,000 isolated and labeled Amazigh handwritten characters written by 60 different writers. This is the only dataset in Amazigh which has contained handwritten characters so far. The AMHCD's database key main purpose to provide the training and testing set for Amazigh handwriting recognition research. A large part of this database has been used in a recent work on recognition of Amazigh handwritten characters [22]. Currently, we are making our efforts to further enlarge the database; meanwhile, we are looking to expand this database by adding more samples of characters extracted from Amazigh texts.

7. ACKNOWLEDGMENTS

Thanks are due to everyone who participated in the data collection stage. I would also like to thank all member of the IRF-SIC Laboratory of the Ibn Zohr University in Agadir, Morocco. A special thanks goes to my friends Mohamed El Hajji and Mustapha Amrouch for their friendship and their valuable help.

8. REFERENCES

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

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

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

[4] 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.

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

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

[7] S. Alma'adeed, D. Elliman, and C.A. Higgins, "A Data Base for Arabic Handwritten Text Recognition Research", The International Arab Journal of Information Technology, Vol. 1, No. 1, January 2004.

[8] 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.

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

[10] M. Ziaratban, K. Faez, F. Bagheri, "FHT: An Unconstraint Farsi Handwritten Text Database", ICDAR'09, 2009, pp. 281-285.

[11] A. Oulamara, J Duvernoy, "An application of the Hough transform to automatic recognition of Berber characters", Signal Processing, vol. 14, 1988, pp.79-90.

[12] A. Djematen, B. Taconet, A. Zahour, "A Geometrical Method for Printing and Handwritten Berber Character Recognition", ICDAR'97, 1997, p. 564.

[13] Y. Ait ouguengay, M. Taalabi, "Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage", Systèmes intelligents-Théories et applications, 2009.

[14] Y. Es Saady, A. Rachidi, M. El Yassa, D. Mammass, "Reconnaissance Automatique de l'Ecriture Amazighe à base de Ligne Centrale de l'Écriture", 4ème Atelier international sur l'amazighe et les TIC, 2011, IRCAM, Maroc.

[15] M. Amrouch, A. Rachidi, M. Elyassa, D. Mammass, "Handwritten Amazigh Character Recognition Based On Hidden Markov Models", ICGST-GVIP Journal, Vol.10, Issue 5, 2010, pp.11-18.

[16] Y. Es Saady, A. Rachidi, M. El Yassa, D. Mammass, "Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata", ICGST-GVIP Journal, Vol.10, Issue 2, 2010, pp.1-8.

[17] R. El Ayachi, K. Moro, M. Fakir, B. Bouikhalene, "On the Recognition of Tifinaghe Scripts", Journal of Theoretical and Applied Information Technology, Vol.20, No.2, 2010, pp.61-66.

[18] Fatima Boukhris, Abdallah Boumalk, El Houssain El Moujahid, Hamid Souifi, "La Nouvelle Grammaire de l'Amazighe", Centre de l'Aménagement Linguistique, Publications de l'IRCAM, Rabat, 2008.

[19] Tifinagh at: http://fr.wikipedia.org/wiki/Tifinagh

[20] Proposition d'ajout de l'écriture Tifinaghe au répertoire de l'ISO/CEI 10646 (format Unicode), 21/06/2004, CEISIC, IRCAM, Rabat, Maroc.

[21] D. S. Le, G. R.Thoma and H. Wechsler, "Automatic page orientation and skew angle detection for binary document images", Pattern Recognition 27, 1994, pp.1325-1344.

[22] Y. Es Saady, A. Rachidi, M. El Yassa, D. Mammass, "Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character", Accepted by IJAST journal, in press.

Table 2: Some Amazigh handwriting characters samples