Amazigh Characters Automatic Recognition: Overview and Prospects

Ali Rachidi[1], Mustapha Eddahibi[2], Youssef Essaady[3], Mustapha Amrouch[4]
[1, 2] ERTISED Research Team, National School of Business and Management, Ibn Zohr University
BP. 37/S, Salam, Agadir, Morocco
[3] Polydesciplinary Faculty, Ibn Zohr University, Taroudant, Morocco
[4] High School of Technology, Ibn Zohr University, Agadir, Morocco
a.rachidi@uiz.ac.ma, m.eddahibi@uiz.ac.ma
, y.essaady@uiz.ac.ma, m. amrouch@uiz.ac.ma

Abstract-- The design and implementation of Amazigh OCR systems is very crucial for the Amazigh language promotion and development. Till now, there is a lack of such system. Therefore, there is growing interest for Amazigh optical character recognition in research works. Indeed, some research works have led to design of systems improving this situation. In this paper, we describe different Amazigh OCR systems and approaches developed and tested in our research team, showing their characteristics and results. The aim of this description is to achieve a comprehensive summary of the various approaches and systems which could help open up some interesting new prospects.

Keywords—Tifinagh; Offline Character Recognition; HMM; Finite Automata.

1 INTRODUCTION

In the OCR field, several scientific research works were performed for Latin-based, Arabic and other scripts. These works have led to the development of several approaches on automated characters recognition and consequently optical scanner to scan and automatically recognize documents. However, The Amazigh character, called Tifinagh, is rarely dealt with. Some approaches have been suggested for Amazigh characters recognition. These approaches are grouped generally into some broad categories such as statistical approaches [20], [8], neural network based approaches [1], [10], [6], [12] [16], syntactical approaches [13], Hidden Markov Models [2] [3] [4] [5] and dynamic programing based approaches [11]. In this paper, we are going to present a state of the art and a summary presentation and comparison of scientific research works accomplished and published in the field of automated recognition of Amazigh printed or handwritten characters, in our team or elsewhere.

The first part of this article deals with main Amazigh characters databases developed for approaches test and validation. The second part is devoted to the description of various works carried out in the field of automated recognition of Amazigh script. In the third part we present our contributions in this field. Finally, we give a detailed overview and comparison of these systems, specifying the benefits as well as the disadvantages of each system. We also, generate a number of perspectives to be adopted in future works.

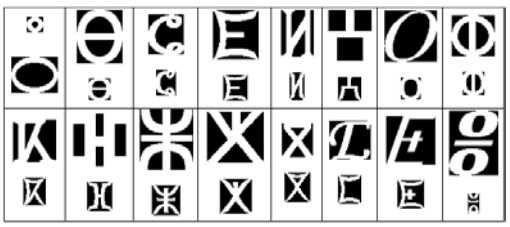
2 MAIN AMAZIGH CHARACTERS DEVELOPED DATABASES

To experiment and validate different developed approaches and systems, it is essential to create Amazigh characters database. Databases of Amazigh characters marked images are unavailable. This field lacks a reference database that provides objective comparisons of various recognition systems. All works published in this field, were tested using local databases that contain a limited number of Amazigh characters. In the next subsections, we describe the two existing characters databases.

2.1 Database of Amazigh spelling patterns

It is a base of patterns of different Amazigh fonts and in various sizes suggested by Ait Ouguengay [1]. It contains a total of 12 character fonts in sizes from 10 to 28 points for each model. Patterns are provided as a set of bitonal images in various sizes. The maximum size is 102x129 pixels, while the minimum size is 19x2 pixels. This disparity can be explained by the fact that the character ya (a) is a small circle. Therefore it is significantly smaller than other characters. In addition to the particular case of the character ya (a), the database is made up of patterns of various Amazigh fonts in various sizes that are not normalized. The table below shows a sample of patterns from this database.

TABLE 1: SAMPLE OF CHARACTERS IN THE AMAZIGH SCRIPT PATTERNS DATABASE



In this database, the way patterns images are created wouldn't enable to renormalize their size into a medium fixed size. Indeed, this can be rather inconvenient particularly because of the close resemblance between the

two characters ya (a) and yar (r). These two characters can be distinguished only by their sizes: The character ya (a) is a small circle, while yar (r) is a bigger circle. In some cases, we will have a real confusion between the two classes of images.

2.2 Handwritten characters database

It is an Amazigh handwritten characters database (A Database for Amazigh Handwritten Character Recognition Research: AMHCD) that we have developed. The database contains 25740 images of Amazigh handwritten characters (Tifinagh) isolated and marked that are produced by 60 writers from different ages, sex and function. The reader can find a complete and detailed description in this database [12].

To date, AMHCD database is little used and explored to evaluate Amazigh script recognition systems [2] [3] [4] [16]. By contrast, it sets itself as the unique and the first database in its kind (handwritten Amazigh script), thanks to its large size and to its availability for academic research.

The table 2 shows a sample of handwritten Amazigh characters. Each character is given in the form of two variations corresponding to two different writers.

TABLE 2: SAMPLE OF HANDWRITTEN AMAZIGH CHARACTERS FROM AMHCD DATABASE [12]

Amazigh printed character	Writer 1	Writer 2	Amazigh printed characters	Writer 1	Writer 2
ⵇ			ⵄ		
ⵅ			ⵃ		
ⵛ			ⵂ		
ⵜ			ⵁ		
ⵍ			ⵀ		
ⵣ			ⵇ		

3 EXISTING AUTOMATED AMAZIGH CHARACTER RECOGNITION APPROACHES

Compared with Latin, Arabic and Chinese, researches on automated Amazigh script recognition have not reached the desired level. As far as we know, very few studies have examined the Amazigh script recognition. In this section, we refer to published works in the field of Amazigh script recognition.

Among the early studies that relates to tifinagh characters recognition, we hold, in the first place, the works of Oulamara published on [20]. The method proposed in that reference is a statistical method based on line segments extractions using Hough transform. Character analysis in the parametric space obtained using Hough transform, allows extracting specific features in association to a reference model that generates the set of alphabet characters. An original character encoding is deducted then used as base to build a reading matrix representing an encoded form of the alphabet. The author has obtained results that seem to be valuable [20] using printed Amazigh characters from a local character database.

Djematene et al. [9] consider that the method published by [20] is not an appropriate technical solution for handwritten Amazigh characters because it produces wrong segmentations. To overcome the difficulty of characters with tilted lines, [9] proposed a statistical method for Amazigh characters recognition based on the position of characteristic points in the rectangle enclosing the character image. After character preprocessing (bidirectional normalization, smoothing, related components extraction), primitives are extracted from each skeleton, as the ends, points, vertices (points of direction changes) and nodes with 3 and 4 branches. Finally, the character representation provides a description form of letters using a predefined encoding. This description encodes positions of characteristic points in the enclosing rectangle. The recognition is to measure the degree of similarity between the elaborated code and reference codes using the metric distance. The results are more or less encouraging for a locally defined characters database despite some errors coming from the preprocessing module.

In [1], Ait Ouguengay proposed an artificial neural network (ANN) for the recognition of Amazigh characters. The neural network used here is a multilayer perceptron with a single hidden layer. The latter was trained on a database that contains patterns of Amazigh script of different fonts and sizes, created locally. The simulation of the neural network was developed using the free JavaNNS software (java neural networks simulator). The used geometrical characteristics are: horizontal and vertical projections, the centers of gravity in x and y, perimeter, area, compactness and second order central moments. According to the author, this approach gives good results on all training patterns. However, the test results are still far from satisfactory because of the test database that is very low compared to the ANN weight to establish.

In [10], the authors propose a Tifinagh script recognition system based on invariant moments and the Walsh transform using dynamic programming. The proposed system contains three main parts: preprocessing, feature extraction and recognition. In the

preprocessing module, the image of the scanned document is cleaned, then, it is segmented into individual characters using histogram techniques. In the feature extraction process, invariant moments and the Walsh coefficients are calculated for segmented characters. Dynamic programming is adopted in the recognition stage. The tests were made on several images of Amazigh writing. According to the authors, the experimental results show that the method of recognition using invariant moments produces better results compared to the method based on the Walsh transform in terms of recognition rate, error rate and time calculation. More recently, the authors have proposed in [11] a multi-layer neural network with the same characteristics previously used. The results found using neural network with one hidden layer are better than those obtained using dynamic programming. In addition, the recognition rate obtained using one hidden layer is higher than that obtained with two or three hidden layers.

In the next section, we present our contributions in terms of design and development of new approaches for automatic recognition of Amazigh printed or handwritten characters. These approached are developed and tested on a large and rich characters database (A Database for Amazigh Handwritten Character Recognition Research :AMHCD) [12].

4 OUR PROPOSED AUTOMATED RECOGNITION SYSTEMS FOR AMAZIGH CHARACTER

4.1 Markov modelling approaches

In [2], we proposed an approach based on hidden Markov models of a type discriminant model DM-HMM approach that focuses on the problems of isolated characters. This type of modeling is widely used in the field of speech recognition. This model is also effective to recognize a form subject to uncertainty and dynamic aspect as Amazigh character. Markov process has implemented specific probabilistic models in order to manage the uncertainty and lack of information that surrounding shapes to recognize. After Amazigh character image preprocessing, the system uses directional primitives in the generation of observation sequences. These observations were obtained using the sliding-window method combined to the standard Hough transform of character images. Sequences of observations obtained are used to train initial HMM models of characters in the learning phase; each model uses samples of his class. Thereafter, we used the Forword classifier to recognize character. Indeed, this approach consists in associating one or several models per class. Thus, recognition is generally done by estimating probabilities of the emission of a series of observations of the form to be recognized by the different models built before. The shape to be recognized is assigned to the class whose model maximizes the probability. This approach is practically used in the case where the number of classes to be recognized is relatively limited, i.e. limited vocabularies such as Amazigh alphabet. However, it becomes more time and memory consuming when the number exceeds a thousand since each class has at least one model of its own.

We evaluated our system's performance with AMHCD the database of Amazigh handwritten isolated characters with two variants. The first adopts the discrete modeling of emission probabilities, whereas the second uses the continuous HMM. The table 3 below shows different recognition rate using the discrete and continuous model based on the number of states of the model and the size of characters base.

TABLE 3: SYSTEM'S RECOGNITION RESULTS IN THE DISCREET AND CONTINUOUS CASES.

Used model Topology	**Discrete HMMs**		**Continuous HMMs**	
	Base size	Recogn rate	Base size	Recogn rate
Modelwith 14 states	2220 Characters	**90.04 %**	-	-
Model mono and 2 Gaussian with 6 states	-	-	25740 Characters	96.21%
Model mono and 2 Gaussian with 10 states	-	-	25740 Characters	96.88%
Model mono and 2 Gaussianwith 14 states	-	-	25740 Characters	97.89%

We believe that the discrete variance errors are mainly: (1) from the discrete model used. At this level, we have recourse to the risk of loss of information; (2) from the used data size to create the basis of the reference models.

To validate this assumption and to deal with this default, we have increased the size of the used database and also we have replaced discrete models by continuous HMMs.

The increase in the training database sizeand modeling probability densities by Gaussian helped to reduce the error rate committed by our system. The overall error rate 9.6% (first experience with discrete HMM) was reduced to 2.11%.The gain in precision is equal to 7.49%. Therefore, the obtained results in the continuous case are better than those obtained in the discrete case.

In [3] and [4], we proposed an alternative system for the recognition of printed Tifinagh characters based on a new approach that exploits the features and morphological characteristics of the Amazigh language.

The brought solution adopts a Markov modeling of the type discriminant path (DP-HMM) optimized by algorithms based on dynamic programming [17] [7]. The approach relies on the proposal for a new list of segments, which consists of a set of basic lines that constitute Amazigh characters. This allows better exploiting of lines redundancy in Amazigh letters plots. Character structure description is based on these elements. Indeed, used features are extracted from character glyphs by its composing segments implicit localization technique. To do so, we used skeletons interest points. In the learning phase, a single HMM global model is built and trained on the elements of vocabulary suggested by structural and geometric primitives. Each path through this lattice represents a sequence of segments, which is a character of the Tifinagh alphabet. The recognition is performed by dynamically decoding optimal path according to the criterion of maximum likelihood.

To validate the proposed system, we have made significant experiments based on data patterns Amazigh script [12]. Several tests were performed to evaluate the recognition rate of the system in terms of: number of states and number of Gaussian mixture. The Table 4 below presents the results of these tests on this basis.

TABLE 4: RECOGNITION RATE ON BDI

States number	3	5
Gaussian number	1-2-3	1-2-3
Recognition rate	**98.41%**	**98.76%**

These results show an error rate of 1.24% using a model with 5 states. We believe that the recognition errors are attributed, on one hand, to methods used for the pre-classification and detection of interest points, and in other hand to the lack of features to be used to better describe each segment. In addition, this low error also comes from the deformation of some characters in some fonts, namely "Tassafut" and "Taromeit" fonts. We note that the number of Gaussians used does not influence the results; while its increase involves a large number of parameters to be calculated. However, the choice of the topology directly influences the results. Therefore, increasing the number of states increases the recognition rate of the system.

4.2 Syntactical approaches

In [13], we presented an automated recognition of printed Amazigh characters based on a syntactic approach using finite automata. After character image preprocessing, applying appropriate algorithms on the skeleton character used to construct representative character string from the Freeman encoding. The reconstructed string is used in the input of a finite automaton that recognizes Amazigh characters. This global automaton was constructed from specific recognition automata of each Amazigh character. We tested our application on the Amazigh printed characters database that we have created. We have obtained encouraging results. Indeed, from 630 read characters, 589 have been recognized, the recognition rate is 93.49%. Table 5 below shows the rates of wrong assigning and wrong rejection. These errors result from the form of some unrecognized characters whose skeleton comprises more non-orthogonal segments. Indeed, the recognition method is based on a vectorization of the character skeleton to be recognized. Therefore, a vectorization error will necessarily result in an error in character description. In fact, the main drawback of this method is the sensitivity of skeleton to noise.

TABLE 5: RECOGNITION ERRORS RATES

Wrong assigning rate	2.28 %
Wrong rejection rate	4.23 %

4.3 Neural approaches

In [10], the authors propose Tifinagh writing recognition system based on the multilayer neural networks with the same characteristics previously used. The results found with a neural network with one hidden layer are better than those obtained with dynamic programming. In addition, the recognition rate obtained using one hidden layer is higher than that obtained with two or three layers.

In [14], [15] and [16], to improve the results of the previous system and have a complete system that recognizes all Amazigh printed and handwritten characters, we have presented a system for automatic recognition of Amazigh text, based on multilayer neural networks. The proposed approach uses statistical primitives based on the position of character's central lines and sliding-window techniques based on works of [19].

Indeed, in this system and in a first step, we used the horizontal central line of the character to extract a set of characteristics of this line based on densities. In a second step, we proposed an improvement to this recognition system by adding other densities features based on character vertical central line in order to exploit the similarity of several Amazigh characters according to character vertical central line. The different variants have been tested and evaluated on two bases: the database of Amazigh script patterns [1] and the database AMHCD for Amazigh handwritten characters. Indeed, we have shown the results of recognition obtained based on the integration of features independent and dependent of character's horizontal central line. Finally, we presented

the results obtained by the improved version and comparing them with those obtained by the first variant of the system.

We also used cross-validation techniques for the evaluation of recognition results. Cross-validation is a method of estimating the reliability of results, based on a sampling technique [18].

Table 6 below shows the results of the proposed system using 10 times cross validation on the basis of patterns of Amazigh script and based on handwriting characters.

TABLE 6: RECOGNITION RESULTS OF IMPROVED SYSTEM BASED ON INTEGRATED FEATURES USING 10 TIMES CROSS-VALIDATION SYSTEM

Integrated features	Patterns database of Amazigh script		Handwritten Amazigh characters database	
	Database size	Recogn Rate	Database size	Recogn Rate
Independent features of the central line	19437 Char	88.68 %	20150 Char	84.49 %
Dependent and independent features of the horizontal central line	19437 Char	**98.49 %**	20150 Char	**92.23 %**
Dependent and independent features of the vertical and horizontal central line	19437 Char	**99.28 %**	20150 Char	**96.32 %**

For the patterns database of Amazigh script, the recognition rate is 88.68% when using only independent features of the horizontal central line and increases to 98.49% when adding features based on horizontal position of the central line. This rate also increases to 99.28% upon the addition of the position features based on the vertical central line.

For the Amazigh handwritten characters database, the rate increases from 84.49% to 92.23% when adding based on the position features of the horizontal central line and increases to 96.32% when adding features based on the position of the vertical central line.

The causes of errors are mainly due to high morphological similarity between certain Amazigh characters and sometimes on different fonts.

Comparing different recognition rates obtained by the system based on integrated features, we see an improvement due to the integration of the features based on the position of the two central lines (vertical and horizontal). This confirms that these features provide significant improvement in the recognition performance.

5 COMPARATIVE SUMMARY

Table 7 summarizes the main features, techniques and results obtained by Amazigh script recognition approaches and systems that we have proposed.

TABLE 7: SUMMARY OF APPROACHES AND SYSTEMS PROPOSED

Authors	Primitives	Models	Results
[2] , [3]	- Directional primitives - Hough Transform	- Discrete and continuous HMM 14 -states - left right topology type. - character learning(MLE) - FORWARD Classification	- 90.4 % in discrete case. - 97.89% in continuous case. - (AMHCD) Database. - Based learning (2/3AMHCD). - Based for test (1/3 AMHCD).
[4], [5]	- Structural primitives (diameter, eccentricity, extent, Center of mass, orientation, length and minimum principal axis moment of order 1 and 2)	- Continuous HMM. - Linear topologies and ergodic - Baum-welsh learning - Viterbi Classification	- 98.76% - Database of patterns of Amazigh script - Based learning (2/3). - Based for test (1/3
[13]	- Freeman coding (skeleton tracking)	-Networks of automata. - Formal grammar	-93.49% According to the data used.

[14]	- statistics primitives - Dependent and independent characteristics of the horizontal and vertical center line - Windows sliding horizontal and vertical	- Multilayer neural networks. - 95 input layer neurons - 31 output layer neurons - layer cover (nb input +nb output) / 2 - Learning of Heb (η=0,2, rate=0,3, itéra=1000)	- 99.28 % - Database of patterns of Amazigh script (19437 characters) - 96.32 % - Database AMHCD (20150 characters)
[10]	The moment invariants and the Walsh transform using dynamic programming	Dynamic programming	Rate of recognition is interesting
[11]	A multilayer neural network with the same features previously used.	Dynamic programming	Rate of recognition is interesting
[16]	Geometric characteristic horizontal and vertical projections, the centers of gravity in x and y, perimeter, area compactness and central moments of order 2.	Neural network	Rate of recognition is interesting
[9]	Ends, the points, vertices (points of direction changes) and nodes 3 and 4 branches.	Statistical method to measure likelihood degree	Results more or less interesting

After having studied and compared different approaches, we find that the best and efficient approach in computing time, memory and recognition rate is statistical primitives and sliding-windows based approach in the preprocessing phase and neural networks in the classification phase [16].

6 CONCLUSION AND PROSPECTS

We have presented in this paper works dealing with automatic recognition of Amazigh characters. The objective is to summarize works done on this subject. These approaches have a number of limitations due both to the preprocessing module and the characteristics recorded in the learning phase. In addition, characters database used in the tests are sometimes weak and sometimes non-standard. Consequently, future research works should make improvements on these approaches on one hand and on the other hand develop other systems that meet expectations. Among our future works, we will develop hybrid systems that use different types of primitives by combining these approaches in the processing phase. This will benefit a priori of advantages of each approach while avoiding main disadvantages.

REFERENCES

[1] Y. Ait Ouguengay, M. Taalabi, Elaboration d'un réseau de neuronesartificiels pour la reconnaissance optique de la graphieamazighe: Phase d'apprentissage, Systèmesintelligents-Théories et applications, Paris :Europia, cop. (impr. au Maroc), ISBN-102909285553, Avril 2009.

[2] M. Amrouch, A. Rachidi, M. El Yassa, D. Mammass, "Handwritten Amazigh Character Recognition Based On Hidden Markov Models", ICGST-GVIP Journal, vol.10, Issue 5, pp.11-18, December 2010.

[3] M. Amrouch, Y. EsSaady, A. Rachidi, M. El Yassa, D. Mammass "Handwritten Amazigh Character Recognition System Based on Continuous HMMs and Directional Features", IJMER journal, Vol.2, Issue 2, pp.436-441, Mar.-Apr. 2012.

[4] M. Amrouch, Y. EsSaady, A. Rachidi, M. El Yassa, D. Mammass "A New Approach Based on Strokes for Printed Tifinagh Character Recognition Using Discriminating Path-HMM ", IRECOS journal, Vol.7, N°.2, Mars 2012.

[5] M. Amrouch, Y. EsSaady, A. Rachidi, M. El Yassa, D. Mammass "A Novel Feature Set for Recognition of Printed Amazigh Text Using Maximum Deviation and HMM", IJCA journal, Vol.44, N°.12, pp.23-30, April 2012.

[6] B.Bouikhalene, M. FAKIR, S.Safi Et B.El Kessab, Reconnaissance Des Caracteres Tifinaghe Par L'utilisation Des Réseaux De Neurones Multicouches, SITACAM 2009, Agadir, Morroco.

[7] R.G. Casey and E. Lecolinet.A survey of methods and strategies in character segmentation.Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol.18,N°.7, pp.690–706, Jul 1996.

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

[9] A. Djematen, B. Taconet, A. Zahour: "Uneméthodestatistique pour la reconnaissance de caractèresberbèresmanuscrits", CIFED'98, pp.170-178, 1998.

[10] R. El Ayachi, K. Moro, M. Fakir, B. Bouikhalene, "On the Recognition of Tifinaghe Scripts", Journal of Theoretical and Applied Information Technology, vol.20 (2), pp.61-66, 2010.

[11] R. EL Ayachi, M. Fakir and B. Bouikhalene, "Recognition of Tifinaghe Characters Using a Multilayer Neural Network", International Journal Of Image Processing (IJIP), vol. 5, Issue 2, 2011.

[12] Youssef EsSaady, Ali Rachidi, Mostafa El Yassa and DrissMammass. AMHCD: A Database for Amazigh Handwritten Character Recognition Research. International Journal of Computer Applications , Vol.27, N°.4, pp:44-48, August 2011.Published by Foundation of Computer Science, New York, USA.

[13] Y. EsSaady, 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, pp.1-8, 2010.

[14] Y. EsSaady, A. Rachidi, M. El Yassa and D. Mammass, "AMHCD:

A Database for Amazigh Handwritten Character Recognition Research", International Journal of Computer Applications, vol.27 (4), pp.44-48, published by Foundation of Computer Science, New York, August 2011.

[15] Y. EsSaady, A. Rachidi, M. El Yassa, D. Mammass, "Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character", International Journal of Advanced Science and Technology, vol.33, pp.33-50, August, 2011.

[16] Y. Essady, M. Amrouch, A. Rachidi, M. El Yassa, D. Mammass, Handwritten Tifinagh character Recognition using Baselines Detection Features, International Journal of Scientific & Engineering Resesarch (IJESR), Volume 5, Issue 4, ISSN 2229-5518, April-2014

[17] H.J. Kim, J.W. Jung, and S.K. Kim.On-line Chinese character recognition using ARTbased stroke classification.Pattern Recognition Letters, Vol.17, N°.12, pp.1311–1322, 1996.

[18] R. Kohavi, "A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[19] B. Ly Van, "Réalisation d'un Système de Vérification de Signature Manuscrite En-ligneIndépendant de la Plateformed'Acquisition", Thèse de doctorat de l'Institut National des Télécommunications,Décembre 2005.

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