

Handwritten Tifinagh Character Recognition Using Baselines Detection Features

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

y.essaady@uiz.ac.ma, amrouch_mustapha@yahoo.fr, rachidi.ali@menara.ma, melyass@gmail.com, mammass@uiz.ac.ma

Abstract—We present in this paper a system of Amazigh handwriting recognition based on horizontal and vertical baselines detection features. After the character image preprocessing, the horizontal and vertical baselines of the character are estimated. Parameters such as baselines are used to derive a subset of baseline dependent features. The symmetry of the amazigh characters with horizontal and vertical baselines is better taken into account. Thus, these features are related to the densities of pixels and are extracted on binary images of characters using the sliding window technique. Finally, a multilayer perceptron is used for character classification. Experimental results using AMHCD database demonstrate the efficiency of the proposed system.

Index Terms — Optical Character Recognition; Amazigh; Tifinagh; Handwritten Recognition; Baseline Detection; MLP

1 INTRODUCTION

Through recent advances in computing power, many techniques of handwriting recognition have also been improved and refined, especially for latin, chinese and arabic scripts [1], [2], [3], [4], [5], [6]. However, the variability inherent in the nature of the handwriting made this area a very active of research. Thus, in recent years, with the growth of means of communication, other alphabets, such as the Tifinagh alphabet of the Amazigh language have integrated the information systems. This has led to the emergence of other types of documents where writing is not yet processed and therefore more difficult to recognize. Text handwritten recognition of such documents requires techniques with more specific treatments.

The Tifinagh is the writing system of the Amazigh language. An older version of Tifinagh was more widely used in 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 [7]. 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 Sahara. The Figure 1 shows a picture of an old Tifinagh script found in site of rock carvings near from Intedeni Essouk in Mali [8].

The Amazigh alphabet which is called "Tifinagh-IRCAM", adopted by the Royal Institute of the Amazigh Culture, was officially recognized by the International Organization of Standardization (ISO) as the basic multilingual plan [9]. The Figure 2 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: X^w (g^w) and K^w (k^w).



Fig. 1. Old tifinagh script, site of rock carvings near from Intédeni Essouk Mali [8]

o	⊖	⊙	⊕	⊘	○	⊚	⊛	
a	b	e	h	u	r	ɾ	s	
⊜	⊝	⊞	⊟	⊠	⊡	⊢	⊣	
ʃ	c	z	ẓ	d	ɖ	f	h	
⊤	⊥	⊦	⊧	⊨	⊩	⊪	⊫	
ε	x	q	i	j	l	m	n	γ
⊬	⊭	⊮	⊯	⊰	⊱	⊲	⊳	
t	ɾ	w	y	g	g ^w	k	k ^w	

Fig. 2. Tifinagh characters adopted by the Royal Institute of the Amazigh Culture with their correspondents in Latin characters

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. In addition, one of the important characteristics of the Tifinagh characters is that the majority of characters have the horizontal or vertical centerline as an axis of symmetry. We present below the Amazigh characters that have the centerlines (horizontal and vertical) as the axes of symmetry.

- Characters which have symmetry orthogonal to the horizontal centerline of the character:

ⵉ, ⵏ, ⵓ, ⵚ, ⵛ, ⵜ, ⵝ, ⵞ, ⵟ, ⵠ, ⵡ, ⵢ, ⵣ, ⵤ, ⵥ, ⵦ, ⵧ, ⵨

- Characters which have symmetry orthogonal to the vertical centerline of the character:

ⵉ, ⵏ, ⵓ, ⵚ, ⵛ, ⵜ, ⵝ, ⵞ, ⵟ, ⵠ, ⵡ, ⵢ, ⵣ, ⵤ, ⵥ, ⵦ, ⵧ, ⵨, ⵩, ⵫, ⵬, ⵭, ⵮, ⵯ, ⵰, ⵱, ⵲, ⵳, ⵴, ⵵, ⵶, ⵷, ⵸, ⵹, ⵺

The recognition of Amazigh handwritten characters has become an active research area because of its potentials and various applications. Recently, some efforts have been reported in literature for Amazigh characters recognition [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. In our previous works, we had proposed a method that contributed to increase the performance of Amazigh characters and handwriting recognition [17], [19]. This approach based on horizontal and vertical centerline of character. After the pre-processing step, the text is segmented into lines and then into characters using the analysis techniques histogram of horizontal and vertical projections. The positions of the centerlines of character are used to derive a subset of baseline independent and dependent features. A subset of the features is related to the position of the horizontal and vertical centerlines to take into account the association of the majority of Amazigh characters by these lines. These features are related to the densities of pixels are extracted on binary images of characters using the sliding window technique. The system showed good performance on two bases of the Amazigh characters: on a printed database of Amazigh characters [11] and on another one for handwritten characters (AMHCD) [16]. The correct average recognition rate obtained using 10-cross-validation was 99.28 % for the 19437 Amazigh printed characters and 96.32 % for the 20150 Amazigh handwritten characters. The causes of errors are mainly due to the resemblance between certain Amazigh characters. Furthermore, the experimental results have showed a significant improvement in recognition rate when integrating the features dependent on the horizontal and vertical centerlines on the base of printed Amazigh characters. Because, the majority of Amazigh characters have the horizontal or the vertical centerline as an axis of symmetry. Nevertheless, the results based on the handwritten characters are still limited. In fact, the high variability of handwriting influences the symmetry of the characters from the horizontal and vertical centerlines. To overcome these limitations, we will use, in this paper, a varied baseline of the character instead of taking the centerline that will improve the results. The architecture of the proposed system is shown in Figure 3.

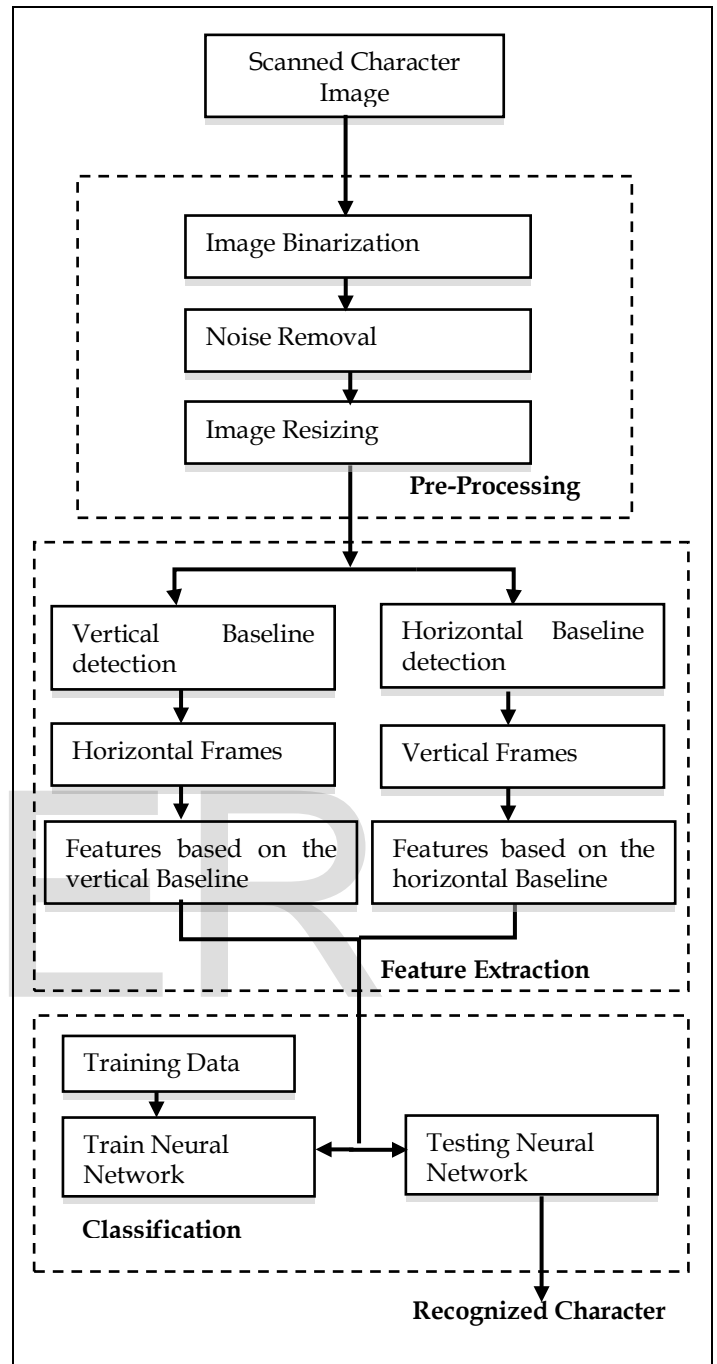


Fig. 3. Architecture of the our System

The paper is organized as follows: Section (2) discusses the need for preprocessing. Section (3) describes features extraction using horizontal and vertical baselines detection. In Section (4), we present the training and recognition steps. Results are given in section (5) and conclusions and future works in section (6).

2 PRE-PROCESSING

Such as was presented in Figure 3, the procedure of pre-processing which refines the scanned input image includes several steps: Binarization, for transforming gray-scale images

in to black and white images, noises removal and image resizing.

We used the Otsu method for binarization [21]. This method of thresholding is performed as a preprocessing step to remove the background noise from the picture prior to extraction of characters and recognition of text. This method performs a statistical analysis of histograms to define a function to be maximized to estimate the threshold.

3 FEATURES EXTRACTION

Before feature extraction, each individual character is uniformly resized into 60×50 pixels. Then, two baselines are extracted in the image of each character; the horizontal baseline and the vertical baseline. After, two different sets of features (45 features based on the vertical baseline and 45 features based on the horizontal baseline giving a total of 90 features) are used in classification and recognition stage.

3.1 Horizontal and vertical Baselines detection

Baseline extraction is generally used for word segmentation into characters or for skew normalization [22], [23], [24], [25], [26]. In this paper, we use horizontal and vertical baselines position to extract baseline dependent features. The horizontal baseline divides the image of the character into two zones: an upper zone that corresponds to the area above the horizontal baseline, and a lower zone which corresponds to the area below the baseline. Our approach is based on the algorithm described in [26] with few alterations. It is based on the projection method and analysis of the maxima and minima of the contour. Figure 4 provides an example of extracted horizontal and vertical baselines of some amazigh handwritten characters.

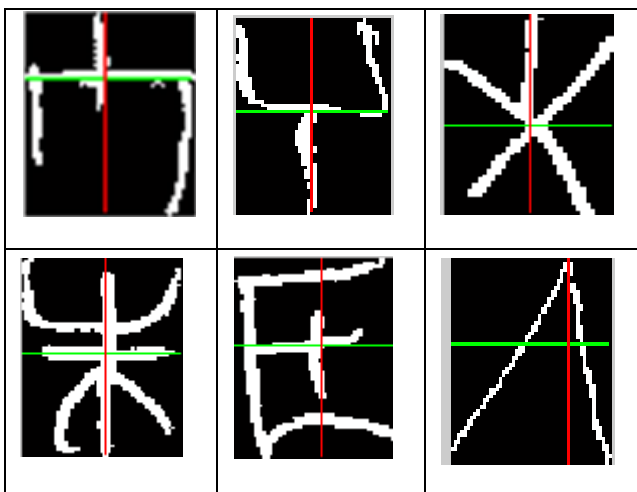


Fig. 4. Horizontal and vertical baselines on some amazigh handwritten characters

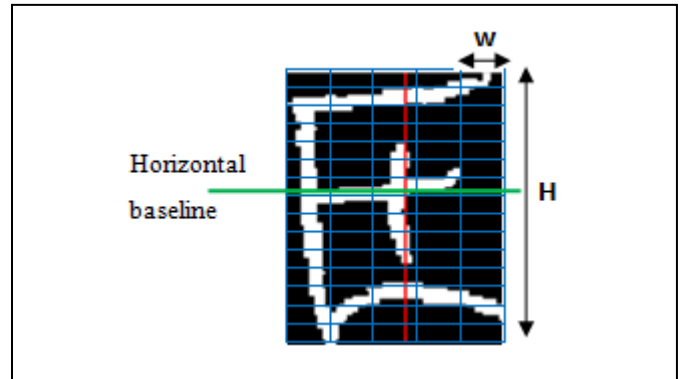
3.2 Features based on the horizontal Baseline

To create the first group of the feature vector, the letter is then scanned from left to right and from top to bottom with a sliding window. The image is divided into vertical frames. The height and width of the frame are constant and are considered

as parameters. The window height varies according to each image. Each frame is divided into cells (Figure 5) where the height cell is fixed (in our experiments to 4 pixels) [22]. This yields to variable number cells in each frame according to the word image height.

In each window, we generate a set of 9 features. These are representative of the densities of pixels.

Fig. 5. Dividing letters into vertical frames



Let H be the height of the frame in an image, h be the fixed height of each cell, and w the width of each frame. The number of cells in each frame n_c is:

$$n_c = \frac{H}{h} \quad (1)$$

Suppose $n(i)$ is the number of black pixels in cell i , $r(j)$ is the number of black pixels in the j^{th} row of a frame, and $b(i)$ is the density level of cell i . If the $n(i)$ of cell i is equal to 0 then we assign 0 to it, otherwise, we assign 1. This procedure is presented in (2).

$$b(i) = 0 \text{ if } n(i) = 0 \text{ else } b(i) = 1 \quad (2)$$

Let HB the position of the horizontal baseline, we consider it the baseline of the letter. For each frame 9 features are extracted. The first feature, f_1 , is the density of foreground pixels (black).

$$f_1 = \frac{1}{H \times w} \sum_{i=1}^{n_c} n(i) \quad (3)$$

The second feature, f_2 , is the number of transitions between two consecutive cells of different density levels of a frame.

$$f_2 = \sum_{i=2}^{n_c} |b(i) - b(i-1)| \quad (4)$$

The third feature, f_3 , is the derivative feature which represents the difference between the gravitational center of black pixels of frame t and its previous frame.

$$f_3 = g(t) - g(t-1) \quad (5)$$

Where g is the position of gravitational center.

$$g = \frac{\sum_{j=1}^H j.r(j)}{\sum_{j=1}^H r(j)} \quad (6)$$

The vertical position of the gravity center in each frame is considered as the fourth feature. This feature is normalized by the height of each frame.

$$f_4 = \frac{g - HB}{H} \quad (7)$$

The fifth and sixth features that represent the density of foreground pixels over and under the horizontal baseline for each frame.

$$f_5 = \frac{1}{H \times w} \sum_{j=HB+1}^H r(j) \quad (8)$$

$$f_6 = \frac{1}{H \times w} \sum_{j=1}^{HB-1} r(j) \quad (9)$$

The seventh and eighth features are similar to the third one. But only those cells that are above the horizontal baseline are considered in, and the cells that are below the horizontal baseline are considered in, f_7 .

$$f_7 = \sum_{i=2}^k |b(i) - b(i-1)| \quad (10)$$

$$f_8 = \sum_{i=k}^{n_c} |b(i) - b(i-1)| \quad (11)$$

The ninth feature, f_9 , is the density of foreground pixels in the horizontal baseline. This feature was added to distinguish between some resembling Amazigh characters.

3.3 Features based on the vertical Baseline

For exploiting the information coming from the detection of the vertical baseline of the character, we generated a new group of the features. To create the second group of the feature vector, the letter image is then scanned from top to bottom and from left to right with a sliding window. The image is divided into horizontal frames. The number of the frame is constant and is considered as one of the system parameters (5 frames in our experiments). Each frame is divided into cells (Figure 6) where the cell height is fixed (in our experiments to 4 pixels). In each window we generate a set of 9 features. These 9 features (f_{10} to f_{18}) are similar to (f_1 to f_9), respectively. We use the same formula used for vertical windows.

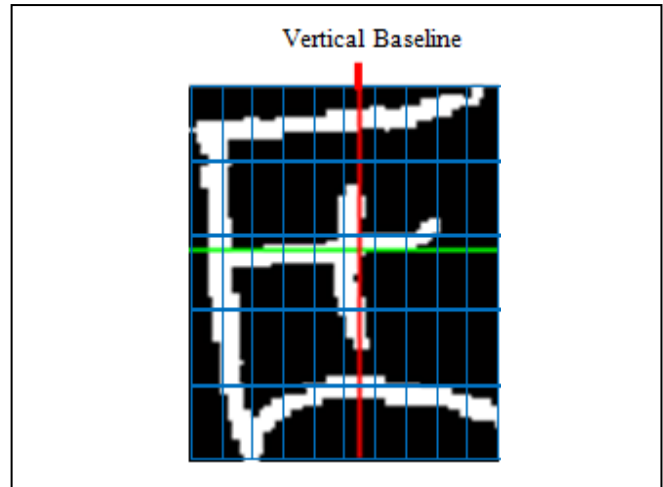


Fig. 5. Dividing letters into horizontal frames

4 TRAINING AND RECOGNITION

In this paper, we choose a multi-layer perceptron architecture using the back propagation with momentum learning scheme. The multilayer perceptron architecture consists of an input layer of information processing nodes, a hidden layer with hidden nodes and an output node which consists of output nodes that usually equal the number of data classes. The Figure 7 presents the Artificial Neural Networks (ANN) Architecture. The number of neuron outputs is 31. The numbers of hidden nodes vary across different applications and this number must be experimentally determined [27], [28]. In our case, we adopted a number of hidden layer neurons equivalent to the number of attributes and classes divided by two.

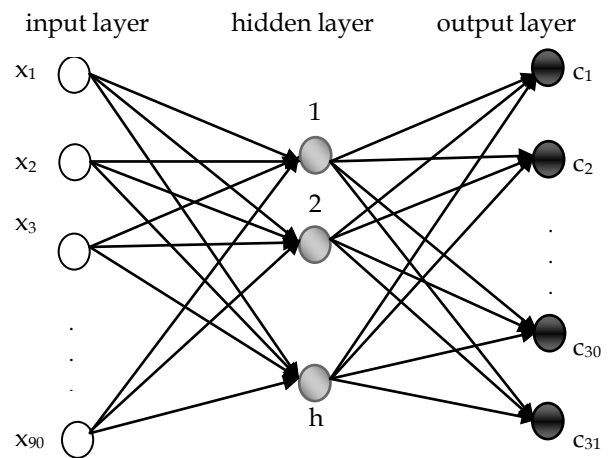


Fig. 6. ANN Architecture (three layers: input, hidden and output)

The network learned on the entire training set using the back propagation method, and was then tested (with the validation set) for its performance. This multilayer perceptron implementation is available in the Weka collection of machine learning algorithms [29].

5 RESULTS AND ANALYSIS

We have tested the proposed system on the base AMHCD that content 24180 handwritten amazigh characters [16]. The tests were performed according to the lines used for the feature extraction (horizontal and vertical baselines or horizontal and vertical centerlines). Also, we have used a 10-fold cross validation scheme for recognition result evaluation [30]. For 10 fold cross validation, 90% of the data was used for training and this performance was tested on the remaining 10%. We divided database in 10 major groups and measured recognition accuracy for each group separately. The recognition rates of all the 10 test subsets of the dataset are averaged to get the recognition result. Table 1 shows the experiments results on the base of handwritten characters (AMHCD) using ten-fold cross-validation.

TABLE 1
 RECOGNITION RESULTS ON THE BASE OF HANDWRITTEN CHARACTERS USING TEN-FOLD CROSS-VALIDATION

Features integrated	Recognition rate %
Features dependent and independent on the horizontal and vertical centerlines	94.62 %
Features dependent and independent on the horizontal and vertical baselines	94.96 %

The recognition rate is 94.62% when integrating the features based on the position of the horizontal and vertical centerlines and increases to 94.96% when replacing the centrelines by the baselines. This demonstrates that the features based on the position of baselines offer a significant improvement in the recognition performance. The most errors are mainly due to the resemblance between some Amazigh characters.

6 CONCLUSION AND PERSPECTIVES

In this paper, we have presented a system for automatic recognition of handwriting Amazigh character based on the position of the vertical and horizontal baselines of each character. Several features have been studied and compared. The importance of using the position of the baselines in the image of the character has been proved. The extracted features are based on the density of pixels derived from a sliding window. The developed system was tested on the database AMHCD of handwriting Amazigh character. These results show a significant improvement in recognition rate when integrating the features dependent on the vertical and horizontal baselines. In future work, we improved the method of baselines detection. In addition, we will add other features that improve the results for some characters, such as information on the possible inclination of writing using the windows inclined.

REFERENCES

[1] R. M. Bozinovic and S. N. Srihari, "Off-line cursive script word recognition", *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 11, no. 1, 1989, pp. 68-83.
 [2] I. Bazzi, R. Schwatz and J. Makhoul: "An Omni font Open-Vocabulary OCR System for English and Arabic. *IEEE Transactions on pattern Analysis and Machine Intelligence*". vol. 21, no 6. 1999, pp. 495-504.

[3] Z. Zhang, L. Jin, K. Ding, X. Gao: Character-SIFT: "A Novel Feature for Offline Handwritten Chinese Character Recognition", 10th International Conference on Document Analysis and Recognition, 2009, pp. 763-767.
 [4] Pooja Agrawal, M. Hanmandlu, Brejesh Lall, "Coarse Classification of Handwritten Hindi Characters", *International Journal of Advanced Science and Technology*, Vol. 10, September 2009, pp. 43-54.
 [5] Elsayed Radwan, "Hybrid of Rough Neural Networks for Arabic/Farsi Handwriting Recognition", (*IJARAI*) *International Journal of Advanced Research in Artificial Intelligence*, Vol. 2, No. 2, 2013, pp. 39-47.
 [6] Sreeraj.M, Sumam Mary Idicula, An Online Character Recognition System to Convert Grantha Script to Malayalam, (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, Vol. 3, No. 7, 2012, pp. 67-72.
 [7] M. Ameur, A. Bouhjar, F. Boukhris, A. Boukouss, A. Boumalk, M. Elmedlaoui, E. Iazzi, "Graphie et orthographe de l'Amazighe", *Publications de Institut Royal de la Culture Amazighe*, Rabat Maroc, 2006.
 [8] <http://fr.wikipedia.org/wiki/Tifinagh>
 [9] L. Zenkour, "L'écriture Amazighe Tifinaghe et Unicode", in *Etudes et documents berbères*. Paris (France). n° 22, 2004, pp. 175-192.
 [10] A. Djematen, B. Taconet, A. Zahour, "A Geometrical Method for Printing and Handwritten Berber Character Recognition", *ICDAR'97*, 1997, p. 564.
 [11] 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.
 [12] R. El Yachi and M.Fakir, "Recognition of Tifinaghe Characters using Neural Network", *International Conference on Multimedia Computing and Systems*, Actes de ICMCS'09, Ouarzazate, Maroc, 2009.
 [13] M. Amrouch, Y. Es-Saady, A. Rachidi, M. Elyassa, D. Mammass, "Printed Amazigh Character Recognition by a Hybrid Approach Based on Hidden Markov Models and the Hough Transform", *International Conference on Multimedia Computing and Systems*, Actes de ICMCS'09, Ouarzazate, Maroc, 2009.
 [14] 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.
 [15] R. El Yachi, 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.
 [16] Y. Es-Saady, A. Rachidi, M. El Yassa, D. Mammass: AMHCD: A Database for Amazigh Handwritten Character Recognition Research, *International Journal of Computer Applications* (0975 - 8887 IJCA), Volume 27(4), pp.44-49, ISBN: 978-93-80864-53-2, August 2011.
 [17] Y. Es-Saady, A. Rachidi, M. El Yassa, D. Mammass: "Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character", *SERSC International Journal of Advanced Science and Technology* (*IJAST*), Vol. 33, August, 2011, ISSN 2005-4238, pp.33-50.
 [18] 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.
 [19] Y. Es-Saady, Contribution au développement d'approches de reconnaissance automatique de caractères imprimés et manuscrits, de textes et de documents amazighs, Thèse de doctorat, Université Ibn zohr- Agadir, Janvier 2012.
 [20] M. Amrouch, Y. Es-Saady, A. Rachidi, M. El Yassa, D. Mammass, "A New Approach Based On Strokes for Printed Tifinagh Characters Recognition Using the Discriminating Path-HMM", *International Review on Computers and Software* (*IRECOS*), Vol. 7 N. 2, ISSN 1828-6003, March 2012.
 [21] Otsu N. A threshold selection method from gray-level histograms, *IEEE Trans. Sys,Man, Cyber*, vol. 9, pp. 62-66, 1979.
 [22] El-Hajj R, Likforman-Sulem L., Mokbel C. Arabic handwriting recognition using baseline dependent features and Hidden Markov Modeling, *ICDAR 05*,

Seoul, Corée du Sud, 2005.

- [23] Razzak M., Sher M., Hussain S. Locally baseline detection for online Arabic script based languages character recognition, International Journal of the Physical Sciences Vol. 5(7), pp. 955-959, 2010.
- [24] Aida-zade R. and Hasanov Z. Word base line detection in handwritten text recognition systems, International Journal of Electrical and Computer Engineering 4:5 2009, pp. 310-314.
- [25] AL-Shatnawi A., Khairuddin O. Methods of Arabic Language Baseline Detection – The State of Art, IJCSNS International Journal, Vol.8 No.10, pp. 137-143, 2008.
- [26] Samia S. Maddouri, Haikal El Abed, Fadoua Bouafif Samoud, Kaouthar Bouriel, Nouredine Ellouze, Baseline Extraction: Comparison of Six Methods on IFN/ENIT Database, The 11th International Conference on Frontiers in Handwriting Recognition, Held on August 19-21, 2008 at Concordia University, Montréal, Québec, Canada.
- [27] G.P. Zhang. "Neural Networks for Classification: A Survey". IEEE Trans on Systems, Man and Cybernetics - Part C, vol. 30, no. 4, 2000, pp. 641-662.
- [28] Jayanta Kumar Basu, Debnath Bhattacharyya, Tai-hoon Kim, "Use of Artificial Neural Network in Pattern Recognition", International Journal of Software Engineering and Its Application, Vol. 4, No.2, April 2010, pp.23-34.
- [29] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", San Francisco: Morgan Kaufmann Publishers, second edition, 2005.
- [30] M. Stone, Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society. 36 (1), 111-147, 1974.

IJSER