A Novel Feature Set for Recognition of Printed Amazigh Text using Maximum Deviation and HMM

M. Amrouch, Y. Es-Saady, A. Rachidi, M. El Yassa, D. Mammass
IRF-SIC Laboratory,
University Ibn Zohr, Agadir, Morocco

ABSTRACT
The growing need of Tifinagh characters recognition in several domains in Morocco such as education, telecommunication, etc, has made it a vital area of research. This paper presents a novel set of structural features generated on the Tifinagh character geometry. This set based on a vocabulary that consists of various fundamental strokes, which is generated using the intrinsic morphological characteristics of the Amazigh script. The input text image is undergoing several preprocessing operations: binarization, skew correction, line segmentation, character segmentation and size normalization. Indeed, the obtained isolated characters are first pre-classified into one of two character groups (circular, non-circular) using the Hough Transformation method. Then, each one is described with their points that have maximum deviation and their segments. Thereafter, each segment of the character is transformed into primitives sequence. We use the discriminating path (DP-HMM) recognition system witch operates on proposed vocabulary. Only one model is built and trained on all elements of this vocabulary. Each path through this trellis represents a sequence of segments, i.e. the character of the Tifinagh alphabet. Finally, the recognition is performed by dynamically decoding the optimal path according to the criterion of maximum likelihood. The obtained scores show the robustness of the proposed approach.

Keywords
Tifinagh characters recognition, structural features, DP-HMMs, dynamic programming, maximum deviation.

1. INTRODUCTION
The offline pattern recognition has an undeniable interest in the realization of large applications such as those found in some domains: reading of bank checks, the postal sorting, reading forms, to name but a few. Early studies on handwriting recognition, prior to 1980, attempted to reproduce the results obtained for speech recognition [1-2].

This work lies within the framework of promoting the offline Amazigh writing recognition. Accordingly, it is of significant importance to improve the performance of the already existing systems and develop or propose other approaches. From this perspective, we opted for a method exploiting the morphological characteristics of the Amazigh language using a Markov model optimized by algorithms based on dynamic programming [3]. Several works, in the literature, have recently been carried out by several authors on the Arabic and Latin writing recognition based on hidden Markov models [4-6]. Within the framework of the Tifinagh recognition, a dozen studies have been conducted based on several approaches. The thesis that has recently been presented by Es-Saady [7] is a reference in this field. However, the use of HMMs to these characters is exclusive to our previous contributions [8-10].

In this paper we suggest a new approach to extract a novel set of the structural primitives from the character strokes using a technique of implicit localization of their segments. This approach is based on the character contour points which have the maximum deviation in the features extraction phase. In order to test their performance we propose a recognition system using the discriminating path that is based on dynamic programming operating at the global graph of segments. The recognition task is performed in this graph by finding the best alignment (dynamic programming).

The remaining parts of this paper develop as follows. Section (2) briefly introduces the proposed dictionary whereas section (3) presents an overview of our system. As to section (4), it sheds light on the pre-processing techniques and methods. Section (5) deals with the process of pre-classification of the characters into circular and non-circular groups while section (6) presents the process of extracting features vector from an image of Tifinagh letter. Section (7) is focused on Markov hidden models of segments. The learning and classification steps are presented in section (8). In addition, section (9) is concerned with the numerical results. The paper finally concludes with an analysis of the results and a description of future work.

2. THE SUGGESTED DICTIONARY
The analysis of the morphology of the alphabet Tifinagh reveals some interesting features particularly the redundancy of the horizontals, verticals and diagonal segments in the majority of letters as well as the redundancy of circular forms that differ only by the presence and the position of a line or point (see figure 1).

Figure 1: The basic segments observed on the character "O", "D", "Z", "S", "A", "E" and "C"

Relying on these characteristics, we proposed a new dictionary that contains 10 basic strokes (see table1). This vocabulary has been constructed automatically from the database of printed Amazigh patterns [11].

Table 1: Dictionary of basic strokes

S1	S2	S3	S4	S5
S6	S7	S8	S9	S10

The Tifinagh characters are theoretically and visually composed of elementary segments and arcs. The use of the afore-mentioned basic strokes to describe the structure of a Tifinagh character is a natural process. Therefore, all Amazigh characters can thus uniquely and completely be described by the list of segments that compose them (see figure 2).

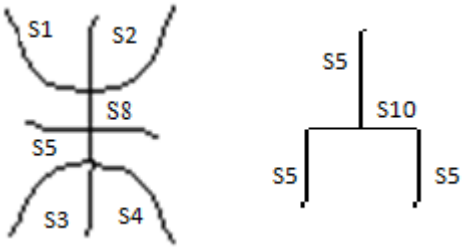


Figure 2: Decomposition of characters « z » and « o » into basic strokes

The explicit segmentation of printed characters in these basic strokes is a very complex problem. Moreover, we note the stability of these strokes in the shape of the Tifinagh characters regardless of the writing style. With regard to this conclusion, we suggest directly considering these strokes as indivisible entities of our Markov modeling.

3. PROPOSED SYSTEM: AN OVERVIEW

The Tifinagh character recognition system we developed is based on a simplified architecture (see figure 3) which deals with the analytic recognition with implicit segmentation and local learning at the level of strokes, through which the system operates in two crucial phases: learning and recognition. Each one is composed of the following steps: preprocessing, normalization, pre-classification and features extraction.

4. PREPROCESSING

In order to refine the scanned input image, it is necessary to perform a set of preprocessing on it. Practically, this set includes several steps for transforming gray-scale images into black and white images, skew correction performed to align the input paper document with the coordinate system of the scanner and segmentation into isolated characters and some operations are used of the size normalization.

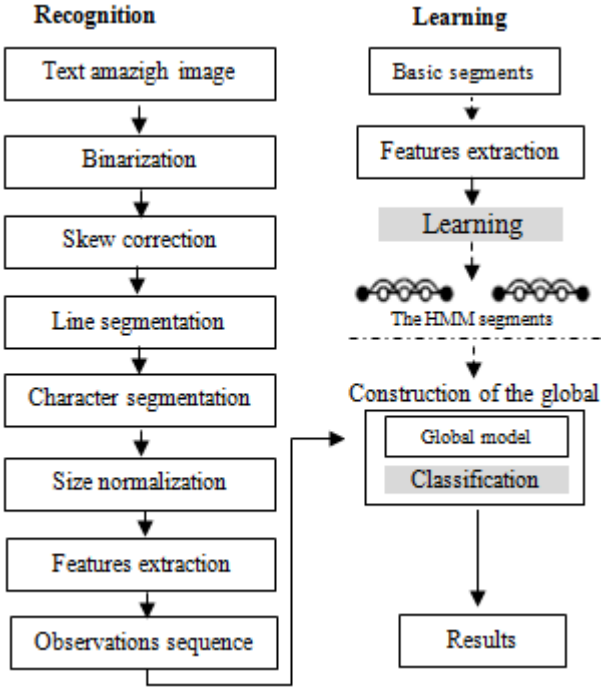
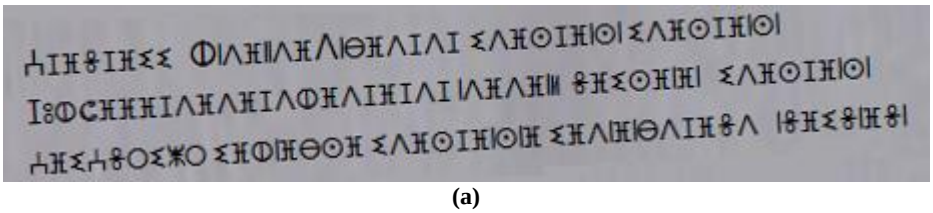


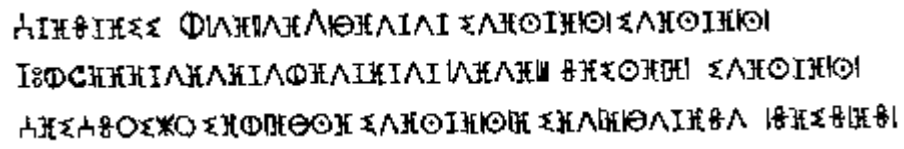
Figure 3: Synopsis of the developed system

Firstly, we used the Otsu method for binarization [12]. 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.

Thereafter, Skew correction methods are used to align the paper document with the coordinate system of the scanner. Main approaches for skew detection include line correlation [13], projection profiles [14], Hough transform [15], etc. For this purpose, two steps are applied. First, the skew angle is estimated. Second, the input image is rotated by the estimated skew angle. In this paper, we use the Hough transform to estimate a skew angle θs and to rotate the image by θs in the opposite direction. Figure 4 (a) shows a sample input text Amazigh image and figure 4 (b) shows the binarized image after the thresholding step using Otsu's method and skew correction algorithm.



(a)

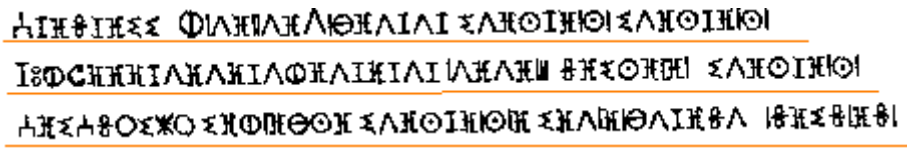


(b)

Figure 4: (a): Text Amazigh's original image; (b): Skew correction and binarization (threshold= 0.8196);

Once the image of the text is cleaned, the text is segmented into lines. This is used to divide text of document into individual lines for further preprocessing. For this, we used analysis techniques of horizontal projection histogram of the pixels in order to distinguish areas of high density (lines) of low-density areas (the spaces between the lines) (see figure 5) [16]. These techniques were often used to extract lines in printed texts, which do not show much variability in the spatial arrangement of related entities such as writing Amazigh. Thereafter, each line obtained is decomposed into characters. Indeed, The Amazigh writing is not cursive, which facilitates the segmenting of a text line of characters. We used the vertical projection histogram to segment each text line of characters. Figure 6 shows a text line, the vertical histogram and the result of segmentation into characters.

Once the text is segmented into isolated characters, the characters' images are converted into images of the same size (48 * 48) using a size normalization algorithm [17].



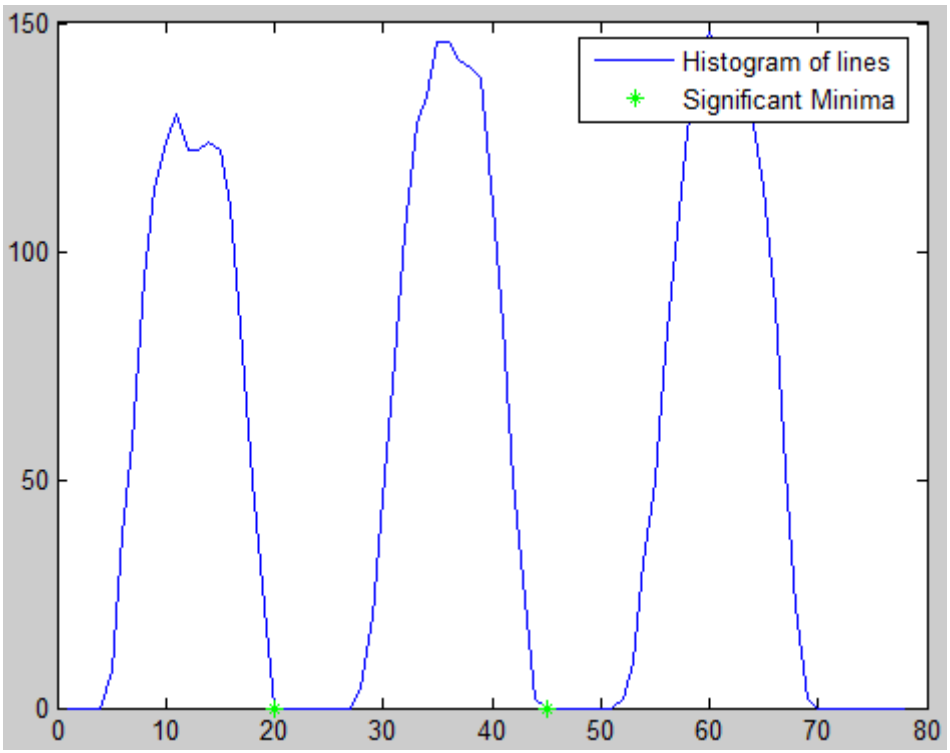


Fig 5: Image text and horizontal projection histograms corresponding

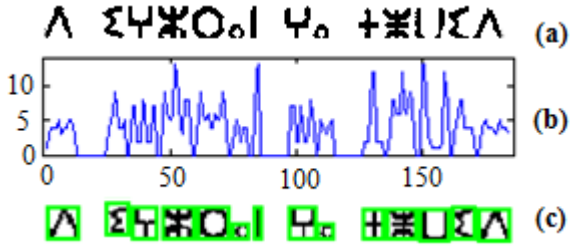


Fig 6: (a) A text line, (b) the vertical histogram, (c) the result of segmentation

5. CHARACTERS' PRE-CLASSIFICATION

This method is based on the feature points of the skeleton. Yet, some characters do not possess them, particularly those containing only circular shapes. Therefore, the main idea is to pre-classify images of characters into two groups: circular

forms (A, r, E, H, U, e, B) and non-circular shapes (m, o, d, p, n, t). With this issue, we use the Hough transformation, which is the most usual effective one [18]. It is based mainly on the implementation of the projection of the image pixels in the Hough space.

Algorithmically, it is simply to browse the image pixels and the elements of the accumulator, and then perform a simple calculation from the equations of circles searched (*R1* and *R2*)

With the radius fixed *R1=Height/2* which corresponds to the characters (Ë, Ã, b, r, s, h) and the radius *R2=Height/4* for the characters (A, e, c, u).

Thus, the values in the accumulator will be the first indications of the presence of circles in the image. However, these values are not sufficient to describe the interest circles for further treatment. Indeed a simple thresholding can then preserve only the vector of parameters whose values are high in the accumulator. These values are actually the searched circles (see figure 7).

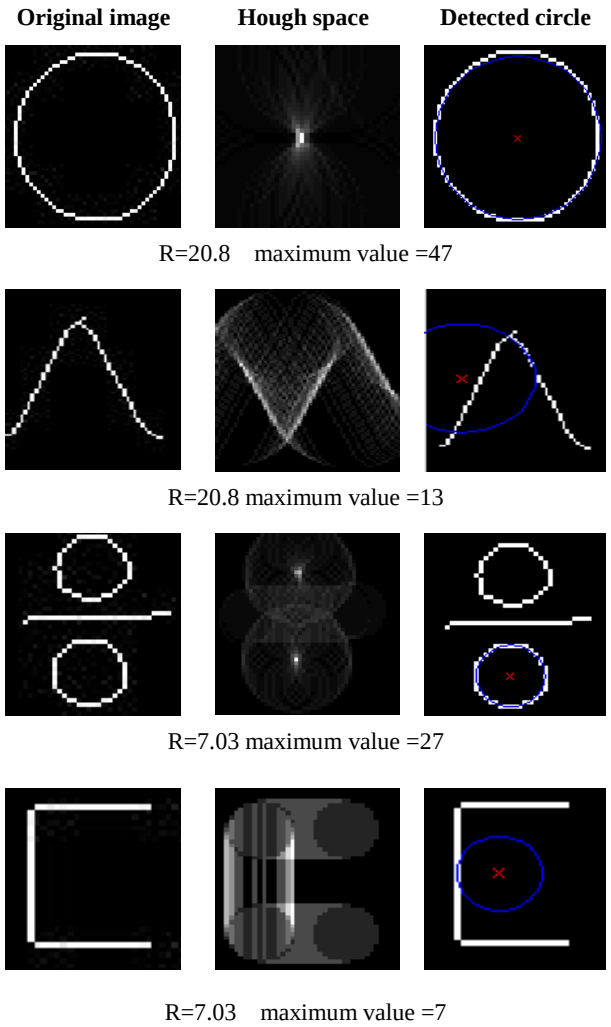


Figure 7: Circles detection by the Hough transform

Consequently, we obtained two classes for the following characters:
– *GC* = {Ë, Ã, b, r, s, h, A, e, c, u};
– GNC = {Z, t, y, i, o, p, d, f, l, m, w, v, n, Ç, Ï, Ä, x, j}.

We continued the filtering process for the *GC* group so as to partition the symbols according to the number N_{cc} of the connected components that compose it. This measure is summarized by the diagram below:

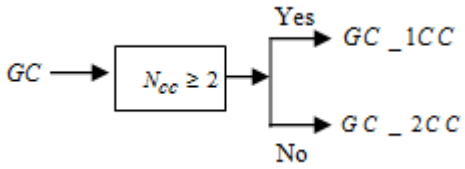


Figure 8: Connected component algorithm

Finally, we obtained the characters' groups:
– *GC_1CC* = {Ë, Ã, b, r, h, A, c };
– *GC_2CC* = {e, u, s};
– *GNC* = {Z, t, y, i, o, p, d, f, l, m, w, v, n, Ç, Ï, Ä, x, j}.

Next, each class has undergone a treatment appropriate to its morphology before being submitted to the common module extraction.

The treatments that have been done were an extraction of the connected components for the character group *GC_2CC*, image character decomposition into 4 parts by their axes central vertical and horizontal for character groups *GC_1CC* and *GC_2CC* (see figure 9).

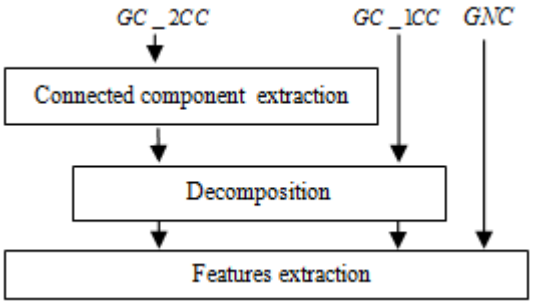


Figure 9: Diagram of pre-classification

6. FEATURES EXTRACTION

This step allows us to represent the input data (characters) into the primitives of a vector fixed size. These primitives must be relevant and discriminating. The features extraction is a crucial and critical step in the recognition systems. In fact, a poor choice of primitives influences negatively and significantly the results even if we use a high performance classifier.

Recently, numerous works have been devoted to research on developing methods for extracting primitives [19][20]. In the literature, the primitives are classified broadly into two categories: structural and statistic primitives.

In our system, we opted for the structural characteristics that describe the topological and geometrical properties of the Amazigh characters, often extracted directly from the skeleton track which together represent:
– The number of holes (Euler Number);
– The minimal and maximal length of the strokes;
– Surfaces, diameters and perimeters;

– The line segments and their attributes (position, center of mass, orientation ...);
– The arcs, loops, and concavities;
– The measurement of curvatures and main directions;
– The measurement of eccentricity, solidity and extent;
– The intersections, junctions, angularity, points and end points.
– The first and the second moment of Hu [21].

In order to count some of these parameters, we proceeded to analyzing the contour of each character for the location of its points which has maximum deviations [22]. The maximum deviation between two points $(x_1\ y_1)$ and $(x_2\ y_2)$ of the contour is given by the equation:

$x\ (y_2 - y_1) + y(x_2 - x_1) + y_2x_1 - y_1x_2 = 0$ *(i.e.* $ax + by + c = 0$*)*

We determine the position and value of the the maximum deviation based on a tolerance threshold empirically determined (threshold = [2 … 4]). If the maximum deviation exceeds allowable tolerance, the contour is reduced in this point and the algorithm continues. In this way each contour is decomposed into line segments (see figure 10).

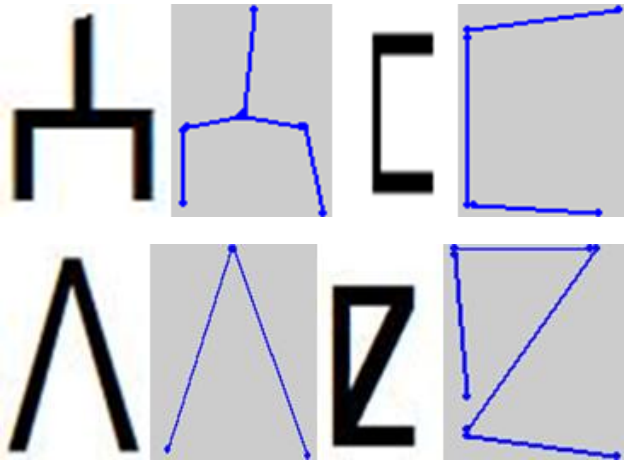


Figure 10: Examples of the localization segments by the maximum deviation

Once the points X_{md} with the maximum deviation are determined on the character contour, we start to extract the primitives' vector from the skeleton, using a contour-tracking algorithm that we proposed. We have fixed a direction of scan from top to bottom and from left to right, the algorithm works well:
We start from a first point X_{md} encountered on the image to track the line until it reaches another point X_{md}, when we reach a junction point X_{md}. In this case, the algorithm proceeds to storing these items one by one in a stack and then follows one of the branches in the direction that we have previously determined until the last point X_{md}. Then, we go back to items stacked to treat the other branches.

A point X_{md} is defined by its coordinates $(x_{md}\ y_{md})$. Each segment is represented by a sequence of pixels delimited by two points $(x_{md1}\ y_{md1})$ and $(x_{md2}\ y_{md2})$ on the skeleton.

During the contour tracking explained above, we calculate a set of structural indices on each of the different segments (strokes) that compose the character based on the bounding boxes of height $h=|y_{i2}-y_{i1}|$ and of width $l=|x_{i2}-x_{i1}|$. However, in the case of the vertical segments where height is zero *(h=0)* and the horizontal segments where width is zero *(l=0)*, we considered windows with dimensions $h=10$ and $l=10$ respectively (see figure 11).

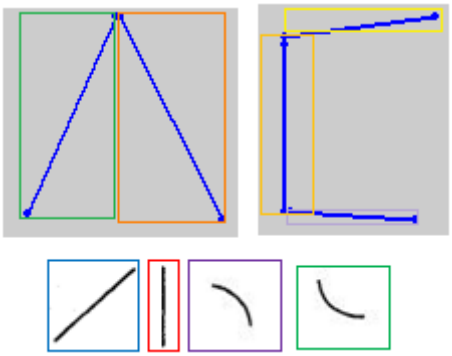


Figure 11: Examples of the bounding boxes

Eventually, for each character we get a vector which includes the primitive sequences generated from each of the segments that compose it.

7. HMMS SEGMENTS

The proposed approach is analytical and based on segments modeling of the dictionary previously presented by Hidden Markov Models (HMMs) [23-24], resulting in a total of 10 models (see figure 12).

The model architecture $\lambda = \{N = 3, M, \Pi, A, B\}$ of a segment is left right (see figure 13) and has the following parameters:

$A = \left\{a_{ij} = p\left(s_j / s_i\right)\right\}$; $\sum_{j=1}^{N} a_{ij} = 1$: The probabilities of states transitions;

$\Pi = \left\{\pi_i = p\left(s_i\right)\right\}$; $\sum_{i=1}^{N} \pi_i = 1$: The vector of the initial probabilities;

$B = \left\{b_i\left(o_k\right) = p\left(o_k / s_i\right)\right\}$; $\sum_{k=1}^{T} b_i\left(o_k\right) = 1$: The probabilities of the observation with $b_j\left(o_k\right)$ are practically estimated by an *M* mixture of multivariate Gaussian distributions:

$$b_j\left(o_t\right) = \sum_{k=1}^{M} C_{jk}\, \left(o_t, \mu_{jk}, \sigma_{jk}\right) \quad 1 \le j \le M, \quad \sum_{k=1}^{M} C_{jk} = 1$$

$$\left(o_t, \mu_{jk}, \sigma_{jk}\right) = \frac{1}{2\pi\sqrt{|\sigma_{jk}|}} \exp\left(1/2\left(o_t - \mu_{jk}\right)\sigma_{jk}^{-1}\left(o_t - \mu_{jk}\right)^T\right)$$

: denotes a normal Gaussian distribution;
C_{jk}: The weight of the k^{th} mixture component;
μ_{jk}: Average vector associated;
σ_{jk}: Covariance matrix associated.

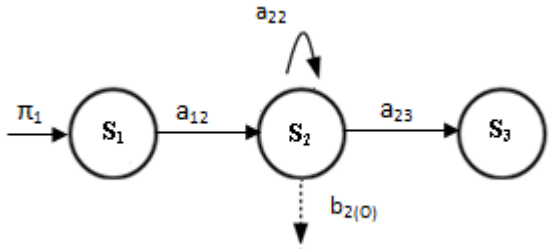


Figure 12: The left-right HMM segments

8. LEARNING AND CLASSIFICATION

The segments models λ_s training is exactly the arduous task of a recognition system. We have sequences of observations obtained from each image segments. We seek to deduce the model that generated them. Once the topologies of the models λ_s were chosen [25][5], learning allows to estimate the probabilities of input, transitions and emissions which best model the base of the samples. This is formally translated by determining the parameters $\lambda_s = \{\pi_s, A_s, B_s\}$ that maximize the likelihood $P(O / \lambda_s)$ of the observations sequence $O=\{o_1, o_2, \ldots o_n\}$.

This is performed by the classical procedure of Baum-Welch [25] which allows re-estimating the HMM of each segment until the likelihood of the generated sequence of the observations is maximal. The best found HMM is saved. Then, all resulting models are concatenated to form the global ergodic model of our system. Each path in this model represents a sequence of segments. Therefore, the recognition of a character is made by finding best path in this graph of segments. We made use of the Viterbi algorithm, which is based on dynamic programming. It allows decoding the best state sequence candidates based on a criterion of maximum likelihood (see figure 13). Practically, it takes the character to be recognized as a sequence of observations $O = \{o_1, o_2, \ldots o_n\}$ extracted from the image and determines the sequence of states $Q=\{q_1,q_2,\ldots q_n\}$ that has the maximum probability of generating O.

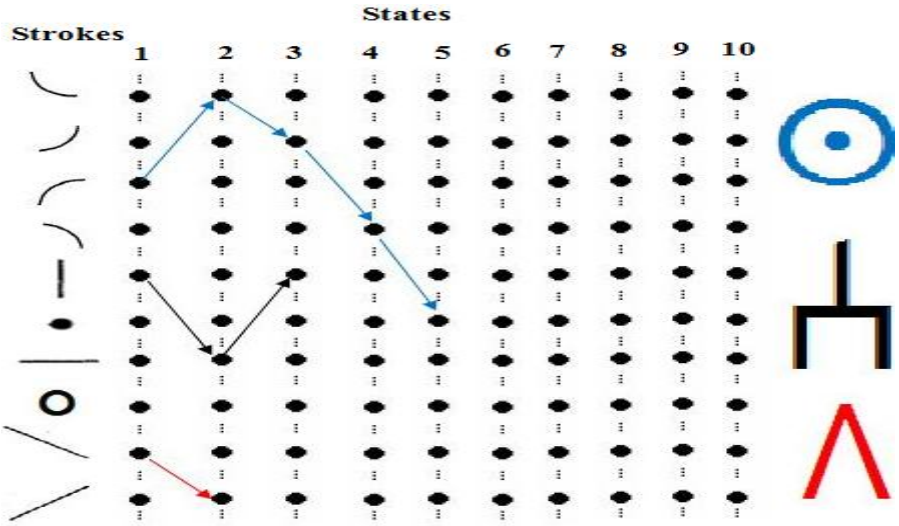


Figure 13: The best paths through the Viterbi network for some characters.

9. EXPERIENCES AND RESULTS

To validate the proposed approach, we performed significant experiments on the database of Amazigh patterns [11]. The database contains the printed Amazigh patterns of different fonts and sizes. It contains a total of 12 fonts and sizes of 10 to 28 points for each model. The patterns are provided as bitonal images with various sizes. We have defined from this database two separate sets of data: a set A (A = 2/3) for learning and a set B (B= 1/3) for testing.

Several tests were performed to evaluate the recognition rate of the system according to the number of states and the number of Gaussian mixture. In addition, the tests done were entirely on the Amazigh patterns database. This database contains 19437 characters of different fonts (i.e. 627 samples × 31 classes). Table 2 shows the obtained results on this database.

Table 2: Recognition rate on database of Amazigh patterns

Number of states	3	5
Number of Gaussian mixtures	1-2-3	1-2-3
Recognition rate	98 , 41%	98 , 76%

Figure 14: Examples of database characters on some fonts

These results show an error rate of 1.14% with a topology model of five states. This fact can be justified by the deformation of certain characters in some fonts, as shown in figure 14 above.

These tests have shown that the combination of the structural's feature and the continuous HMMs are more effective. The best recognition rate recorded was 98,76%. Therefore, the use of the maximum deviation technique for the location of features point, shows a slight increase in the rate recorded over the technique used in our previous approach [26], based on the number of transition from a black pixel to a white (0→1) on it 8-neighborhood. Indeed, we obtained a rate of 98,52% with the same experimental conditions. i.e. A gain of 0.34%.

This experimental results illustrate that this proposed approach is more promising than our previous approach [26] and which gives a rate of 98,52% of good recognition. Finally, we believe that other errors of the recognition are assigned, on the one hand, to the maximum deviation method used and, on the other hand, to the lack of features used to describe each segment. The introduction of other characteristics especially the statistic primitives and the use of continuous HMMs (CHMMs) in combination with other classifiers especially MLP can reduce the error rate of our system. In addition, we will experiment our approach on a database of Amazigh handwritten characters [27].

10. CONCLUSION

In this paper, we proposed a novel technique to exact a set of structural's feature from the character contour based on points that having a maximum deviation. The approach is evaluated by a proposed system of Tifinagh characters recognition in a limited segments' dictionary. The solution provided is based on dynamic programming with continuous HMMs. This modeling has the advantage of being independent of the number of classes of recognition (in terms of memory and speed) since the model is constructed for all classes. The results, which are quite encouraging, have shown that continuous HMMs are more robust. However, the disadvantages of this approach are the detection of the points that have the maximum deviation for features extraction phase seems restrictive for certain fonts of Amazigh writing; To overcome these problems, looking for a suitable integration method for a better estimation of these points is a prospect that we plan for the sake of improving the performance of our system.

11. REFERENCES

[1] Herman Ney and Stephan Ortmanns, Progress in dynamic programming search for LVCSR. Proceedings of the IEEE, 88(8), pp. 1224–1240, August 2000.

[2] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, Spoken language processing, Prentice Hall, 2001.

[3] Nafiz Arica and Fatos T. Yarman-Vural, An overview of character recognition focused on off-line handwriting, IEEE Transactions on Systems, Man and Cybernetics - part C: Applications and Reviews, 31(2), pp.216–233, May 2001.

[4] A. Benouareth Reconnaissance de Mots Arabes Manuscrits par Modèles de Markov Cachés à Durée d'Etat Explicite PhD thesis, Univ Badji Mokhtar, Annaba V, 2007.

[5] X. Dupre, Contributions à la reconnaissance de l'écriture cursive `a l'aide de modèles de Markov caches, PhD thesis, Univ Rene Descartes, Paris V, 2003.

[6] F. Menasri, Contributions à la reconnaissance de l'écriture arabe manuscrite, PhD thesis, Université Descartes, Paris, 2008.

[7] Y. Es-Saady, Contribution au développement d'approches de reconnaissance automatique de caractères imprimés et manuscrits, de textes et de documents amazighes, PhD thesis, Université Ibn zohr, Agadir, 2012.

[8] 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, ICMCS'09, Avril 2009, Ouarzazate, Maroc.

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

[10] M. Amrouch, Y. Es-saady, A. Rachidi, M. El Yassa, D. Mammass, Handwritten Amazigh Character Recognition System Based on Continuous HMMs and Directional Features, International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.2, pp-436-441, Mar-Apr 2012.

[11] A. Ait Ouguengay, Elaboration d'un réseau de neurones artificiel pour la reconnaissance optique de la graphie amazighe, Phase d'apprentissage, SITA'08, 5ème conférence sur les systèmes intelligents : Théories et applications, INPT, Mai 2008, Rabat-Maroc.

[12] N.Otsu. A threshold selection method from grey-level histograms, IEEE Trans. Syst. Man. Cybern., vol.SMC-8, 1978.

[13] H.Yan, "Skew correction of document images using interline cross-correlation", CVGIP: Graphical Models Image Process 55, 1993, 538-543.

[14] T. Pavlidis and J. Zhou, Page segmentation and Classification, Comput. Vision Graphics Image Process. 54, 1992, 484-496.

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

[16] Y. Es Saady, 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.

[17] S.N. Srihari, E.J. Keubert. Integration of handwritten address interpretation technology into the United States postal service remote computer reader system, ICDAR, pages 892–896, 1997.

[18] Henri Maitre, Un panorama de la transformation de Hough, École Nationale Supérieure des Télécommunications, Labo Image, Département Images, Sons et Vidéo, traitement de sgnal, vol. 2, n.4, 1985.

[19] Grandidier F., Sabourin R., Suen C.Y., and Gilloux M., Une nouvelle stratégie pour l'amélioration des jeux de primitives d'un système de reconnaissance de l'écriture, CIFED'2000, pp. 111-120, July 2000, Lyon, France.

[20] Britto, A.S., Sabourin R., Bortolozzi F. and Suen C.Y., "Foreground and Background Information in an HMM-Based Method for Recognition of Isolated Characters and Numeral Strings", 9th IWFHR, pp 371-376, October, 2004, Tokyo, Japan.

[21] M.K. Hu, Pattern recognition by moment invariants, In Proc. IRE, pp. 1428, Sept. 1961.

[22] Perter Kovesi, Invariant Mesures of Image Features from phase Information, rapport thèse, département de psychologie, université de Western Australia, May 1996.

[23] L. R. Rabiner. A Tutorial on Hidden Markov Modelsand Selected Applications in Speech Recognition. Proceedings of the IEEE, vol. 77, no. 2, pages 257–286, 1989.

[24] Rabiner, L., and Juang, B. Fundamentals of speech recognition, Prentice Hall, 1993.

[25] E. Augustin, Reconnaissance de mots manuscrits par systèmes hybrides Réseaux de Neurones et Modèles de Markov Cachés, PhD thesis, Paris, 2001.

[26] 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, accepted by IRECOS Journal, in press, march 2012.

[27] Y. Es Saady, Ali Rachidi, Mostafa El Yassa and Driss Mammass, AMHCD: A Database for Amazigh Handwritten Character Recognition Research. International Journal of Computer Applications 27(4):44-49, New York, USA August 2011.