

Semi-Continuous HMMs with Explicit State Duration Applied to Arabic Handwritten Word Recognition

Abdallah Benouareth, Abdellatif Ennaji, Mokhtar Sellami

► **To cite this version:**

Abdallah Benouareth, Abdellatif Ennaji, Mokhtar Sellami. Semi-Continuous HMMs with Explicit State Duration Applied to Arabic Handwritten Word Recognition. Tenth International Workshop on Frontiers in Handwriting Recognition, Université de Rennes 1, Oct 2006, La Baule (France). inria-00108312

HAL Id: inria-00108312

<https://hal.inria.fr/inria-00108312>

Submitted on 20 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Semi-Continuous HMMs with Explicit State Duration Applied to Arabic Handwritten Word Recognition

Abdallah BENOURETH

Laboratoire de Recherche en
Informatique Institut
d'Informatique- Université
Badji Mokhtar - Annaba- BP,
12- 23000 Sidi Amar, Algérie.
Tel & Fax (213) 38 87 29 04
benouareth@lri-annaba.net

Abdellatif ENNAJI

Laboratoire PSI -FRE 2645
Université de Rouen, INSA de
ROUEN.
abdel.annaji@univ-rouen.fr

Mokhtar SELLAMI

Laboratoire de Recherche en
Informatique Institut
d'Informatique- Université
Badji Mokhtar - Annaba- BP,
12- 23000 Sidi Amar, Algérie.
Tel & Fax (213) 38 87 29 04
sellami@lri-annaba.net

Abstract

The goal of this paper is to describe an off-line segmentation-free Arabic handwritten words recognition system. This system is based on a semi-continuous 1-dimensional hidden Markov models (SCHMMs) with explicit state duration of different kinds (Gauss, Poisson and Gamma). First preprocessing is applied to simplify the feature extraction process, then the word image is analyzed from right-to-left, by using a sliding window approach, in order to extract from it a vectors sequence of statistical and structural features. The extracted sequence is submitted to a SCHMMs classifier based on a likelihood criterion for identifying the word using a modified version of the Viterbi algorithm.

Several experiments were performed using the IFN/ENIT benchmark database, they showed, on the one hand, a considerable improvement in the recognition rate when SCHMMs with explicit state duration of either discrete or continuous distribution are used instead of standard SCHMMs (i.e. with implicit state duration), on the other hand, the Gamma distribution for the state duration, that have given the best recognition rate (90.89 % in top 1), seems more suitable for the SCHMMs based modeling of Arabic handwriting.

Keywords : semi-continuous HMMs, semi-continuous with explicit state duration, modified Viterbi algorithm.

1. Introduction

The term handwriting recognition (HWR) denotes the process of transforming a language, which is presented in its spatial form of graphical marks, into its symbolic representation. Off-line HWR carries out this task after the writing process. Many applications require off-line HWR capabilities such as bank processing, mail sorting, commercial forms-reading, office automation, etc. However, off-line HWR remains a challenging task, in spite of the important boost of research in this field and the latest improvement in recognition methodologies and systems that are very encouraging.

Works on handwriting Arabic recognition, although they are not as advanced as those devoted to other scripts (e.g. Latin), know recently a renewed interest [2] [3]. We

point out, that the fact of directly applying the developed techniques to Latin handwriting recognition on Arabic handwriting is not appropriated, because, Arabic script is based on alphabet and rules different from those of Latin.

Arabic writing, both handwritten and printed is semi-cursive (i.e. the word is a sequence of disjoint connected components called pseudo-words and each pseudo-word is a sequence of completely cursive characters and is written from right to left). Besides, Arabic fertility in diacritics (e.g. dots, Hamza, etc.) makes its recognition process very hard.

Stochastic models, especially hidden Markov models (HMMs) [12], have been successfully applied to the field of off-line HWR in recent years [14]. The main advantage of those models lies in their probabilistic nature, enabling a robust modeling of signals corrupted by noise such as speech or handwriting, and in their capability to integrate efficiently the contextual information at different levels of recognition process (morphological, lexical, syntactical, etc.).

Being aware of those advantages, we have developed an off-line segmentation-free Arabic handwritten words recognition system based on semi-continuous 1-dimensional HMMs (SCHMMs) with explicit state duration of different kinds (Gauss, Poisson and Gamma). We point out that our work makes up, for the off-line Arabic handwritten recognition problem, the first experimentation in using SCHMMs with explicit state duration of discrete and continuous distributions.

After preprocessing intended to simplifying the later steps of the recognition process, firstly, the word image is uniformly or variably segmented, from right to left, into frames. Then, each frame is analyzed in order to be characterized by a 33 size vector combining statistical and structural features. The output of this step is a sequence of feature vectors that are submitted to a SCHMMs classifier for performing word discrimination. The SCHMMs relating to the words of the recognition lexicon are built by concatenating their appropriate characters SCHMMs which are created during an *embedded training stage*. Word identification is carried out by a modified version of the Viterbi algorithm using a maximum likelihood criterion.

Significant experiments have been performed on the IFN/ENIT benchmark database [10], they showed, on the one hand, a considerable improvement in the recognition rate when SCHMMs with explicit state duration of either discrete or continuous distribution are used instead of standard SCHMMs (i.e. with implicit state duration cf. §2), on the other hand, the Gamma distribution for the state duration, that have given the best recognition rate (90.89 % in top 1), seems more appropriate for HMMs based modeling of Arabic handwriting. The resulting performances of varying the distribution kind for the state duration as well as the HMMs parameter choices are discussed.

This paper is organized as follows. Section 2 succinctly introduces the standard SCHMMs and details SCHMMs with different explicit state duration types and their parameters estimation, and presents a modified version of the Viterbi algorithm. Section 3 summarizes by a bloc diagram the system developed. Section 4 details the preprocessing applied to word image. Section 5 describes the step of feature extraction. Section 6 deals with the training and classification process. Section 7 presents and discusses the results achieved by the experiments performed on the IFN/ENIT benchmark database. Finally, a conclusion is given with some perspectives in section 8.

2. Semi-Continuous HMMs and SCHMMs with Explicit State Duration

The semi-continuous Hidden Markov Models (SCHMMs) have been proposed to extend the discrete HMMs by replacing discrete output probability distributions with a combination of original discrete output probability distributions and continuous probability density functions of Gaussian codebook [7]. In the SCHMM each vector quantization (VQ) codeword is regarded as Gaussian probability density. Intuitively, from the discrete HMM point of view the SCHMM tries to smooth the discrete output probabilities with multiple code word candidates in VQ procedure. From the continuous mixture HMM point of view, the SCHMM ties all the continuous output probability densities across each individual HMM to form a shared Gaussian codebook, i.e. a mixture of Gaussian probability densities. With the SCHMM, the codebook and HMM can be jointly re-estimated to achieve an optimal codebook model combination in the sense of maximum likelihood criterion. Such a tying can also substantially reduce the number of free parameters and computational complexity in comparison with the continuous mixture HMM, while maintain practically modeling power of a mixture of a large number of probability density functions.

We distinguish broadly between two semi-continuous HMM types: SCHMM with implicit state duration (i.e.; standard SCHMM) and SCHMM with explicit state duration. Standard SCHMMs do not allow explicit duration modeling (i.e. duration that can spend the model in some state). Indeed, the probability distribution to spend a duration d in the state i (i.e. probability of consecutively observing d symbols in state i), noted $P(d)$,

is always considered as a geometric one with parameter a_{ii} :

$$P(d/q_i) = a_{ii}^{d-1} (1 - a_{ii}) \quad (1)$$

The form of this distribution is exponentially decreasing (i.e. it gets to its maximal value at the minimal duration $d = 1$, and decays exponentially as d increases). Described with one parameter, the distribution can effectively depict only the mean duration. Beyond that it is unable to model any variations in the duration distributions, and hence, its use is not appropriate in the case when the states have some explicit signification. For example, in handwriting they represent the letters or letter fragments, because, in this case, small letters are modeled as being more probable than larger letters. As a results, it is suitable to explicitly model the duration spent in each state. In the presented system, we have used one analytical discrete distribution (i.e., Poisson [13]) and two other continuous (i.e., Normal and Gamma [8]) for modeling state duration.

An SCHMM λ with explicit state duration probability distribution is defined by the following parameters: $A, B, N, p(d)$, and Π that are state transition probability matrix, output probability matrix, a total number of HMM states, a state duration probability vector, and initial state probability vector, respectively.

The model likelihood $P(O/\lambda)$ of an HMM λ with explicit state duration, for a discrete observation sequence O , can be computed by a generalized forward-backward algorithm [8], as follows:

$$P(O/\lambda) = \sum_{i=1}^N \sum_{j=1}^N \sum_{d=1}^t \alpha_{t-d}(i) a_{ij} p_j(d) \prod_{s=t-d+1}^t b_j(o_s) \beta_t(j), \quad (2)$$

where α_i and β_i are, respectively, the partial forward and backward likelihoods.

2.1. Discrete distribution

For the speech recognition purpose, Russel and Moore [13] have used a Poisson distribution for the state duration probability in the HMM. This distribution is defined as follows:

$$p_j(d) = e^{-l_j} \times \frac{(l_j)^d}{d!} \quad (3)$$

The random variable d denotes the spent duration in the state j and follows this distribution, has an expected value l_j representing the one parameter of the Poisson density. This parameter is re-estimated by (4), and it is considered as the expected spent duration in the state j divided by the expected occurrence of this state.

$$l_j = \frac{\sum_{t_0=1}^T \sum_{t_1=t_0}^T \chi_{t_0 t_1}(j) \cdot (t_1 - t_0 + 1)}{\sum_{t_0=1}^T \sum_{t_1=t_0}^T \chi_{t_0 t_1}(j)}, \quad (4)$$

where

$$\chi_{t_0 t_1}(j) = \frac{\sum_{i=1, i \neq j}^N \alpha_{t_0-1}(i) a_{ij} \prod_{s=t_0}^{t_1} b_j(o_s) p_j(t_1 - t_0 + 1) \beta_{t_1}(j)}{P(O/\lambda)}. \quad (5)$$

2.2 Continuous distribution

Levinson [8] has proposed, in the HMMs based speech recognition framework, two continuous distributions for the state duration probability of Gamma and Gaussian kind.

• Gaussian distribution

With this distribution, the state duration probability distribution is defined as follows:

$$p_j(d) = \frac{1}{\sigma_j(2\pi)^{1/2}} \times e^{-\frac{(d-m_j)^2}{2\sigma_j^2}} \quad (6)$$

where, m_j and σ_j are the mean and variance of the Gaussian distribution.

• Gamma distribution

In this case, the state duration density is defined by (7):

$$p_j(d) = \frac{\eta_j^{\nu_j} \nu_j^{\nu_j-1} e^{-\eta_j d}}{\Gamma(\nu_j)} \quad (7)$$

where, the η_j and ν_j are the parameters of the Gamma distribution having a mean $\mu_j = \nu_j \eta_j^{-1}$ and a variance

$$\sigma_j = \nu_j \eta_j^{-2}.$$

The parameters of these continuous distributions are estimated by applying (4) and (8):

$$\bar{\sigma}_j = \frac{\sum_{t_0=1}^T \sum_{t_1=t_0}^T \chi_{t_0, t_1}(j) \cdot (t_1 - t_0 + 1)^2}{\sum_{t_0=1}^T \sum_{t_1=t_0}^T \chi_{t_0, t_1}(j)} - (\bar{\xi}_j)^2, \quad (8)$$

where, $\bar{\mu}_j$ is defined by (4).

2.3. A Modified Version of the Viterbi Algorithm

We propose an extended Viterbi algorithm for sequence decoding in SCHMMs with explicit state duration, which is stated as follows:

(1) **Initialisation** $1 \leq i \leq N$

$$\delta_1(i) = \pi_i b_i(O_1) p_i(1) \quad (9)$$

$$\psi_1(i) = 0 \quad (10)$$

(2) **Recursion** $2 \leq t \leq T$ $1 \leq i \leq N$,

$$\delta_t(i) = \max_{1 \leq \tau \leq t-1} \left\{ \max_{\substack{1 \leq j \leq N \\ i \neq j}} \left\{ \delta_\tau(j) a_{ji} \right\} p_i(t-\tau) \prod_{k=\tau+1}^t b_i(O_k) \right\} \quad (11)$$

$$\psi_t(i) = \arg \max_{1 \leq \tau \leq t-1} \left\{ \arg \max_{\substack{1 \leq j \leq N \\ i \neq j}} \left\{ \delta_\tau(j) a_{ji} \right\} p_i(t-\tau) \prod_{k=\tau+1}^t b_i(O_k) \right\} \quad (12)$$

(3) **Termination**

$$P^* = \max_{1 \leq j \leq N} [\delta_T(j)] \quad (13)$$

$$q_T^* = \arg \max_{1 \leq j \leq N} [\delta_T(j)] \quad (14)$$

$$\tau^* = T - \psi_T(q_T^*, 1) \quad (15)$$

$$q_{T-\tau}^* = q_T^*, \quad 1 \leq \tau < \tau^* \quad (16)$$

(4) **Path backtracking** $T - \tau^* \geq t \geq 1$

$$q_t^* = \psi_{t+\tau^*}^*(q_{t+\tau^*}^*, 2) \quad (17)$$

$$\tau^* = t - \psi_t(q_t^*, 1) \quad (18)$$

$$q_{t-\tau}^* = q_t^*, \quad 1 \leq \tau < \tau^* \quad (19)$$

$$t = t - \tau^* \quad (20)$$

where, $\delta_t(i)$ is the probability of the best state sequence ending in state S_i at time t , but which can be in other state at time $t+1$. The variable $\psi_t(i)$ is an array of size 2 used to memorize the state sequence of the optimal path and the duration of each state in this sequence, i.e., $\psi_t(i,1)$: the duration spent in the state i ; $\psi_t(i,2)$: the state leading to the state i .

3. Developed System Block Diagram

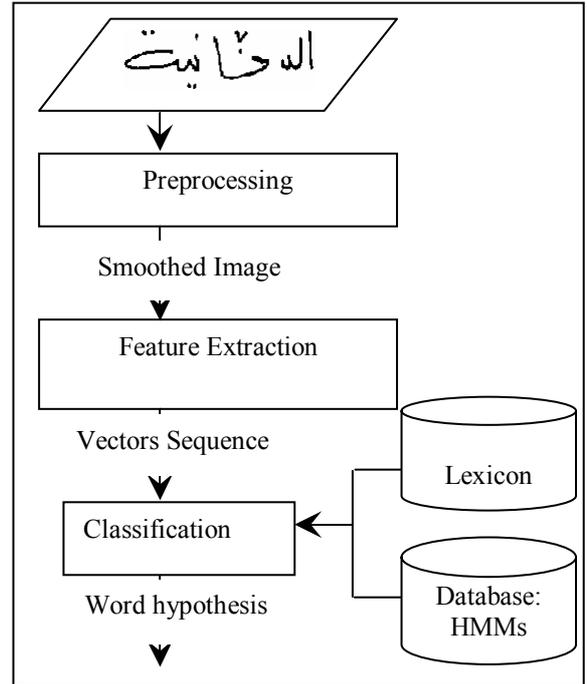


Figure 1. Block diagram of the developed system.

4. Preprocessing

The aim of preprocessing is the removal of all the elements in the word image that are not useful to recognition. Typically, preprocessing consists in some operations such as binarization, smoothing, baseline detection and thinning. Due to fact that we use the

cropped binary word images coming from IFN/ENIT – database [10], binarization has already been done during the database development. A smoothing process was taken to perform noise reduction by using the spatial filter proposed by Amin [1]. The extraction of some features (i.e., ascenders and descenders) requests baseline (i.e., writing line) detection in the word image. For this purpose, the algorithm presented in [6] has been used and has given good results. It is based on a two-step linear regression applied to the local minima points of word contour. Furthermore, we extract the lower and upper baselines of the word image. These baselines split the image into: a middle zone that doesn't contain ascenders and descenders, and 2 zones where ascenders and descenders can be found respectively [5]. The method described in [4], based on the horizontal projection curve that is computed with respect to the horizontal pixels density, has been applied with satisfaction. Thinning is used in order to bring down handwriting style variability and to make straightforward the extraction of some feature such as cusp points, loop, etc. This operation is time consuming, and sometimes its application to Arabic handwriting can remove diacritic points which are relevance primitives for word discrimination. The Pavlidis's algorithm [9] has a tolerate complexity and preserves the diacritic points.

5. Feature Extraction

Our feature extraction process starts by segmenting, from right to left, the word image into many frames. Then, each frame is characterized by a parameter vector. We have adopted two segmentation schemes into frames. The first one is uniform (i.e., all frames have an identical wide), see Figure 2.a. The second is no-uniform (i.e., the frames have not necessary the same wide), see Figure 2.b. In the last scheme the boundary (i.e., the start and the end column) of each frame is based on minima and maxima analysis of the vertical projection histogram. The best frame wide in the uniform segmentation has been empirically fixed to 10. We point out that the non-uniform segmentation gave better performance than the uniform one.

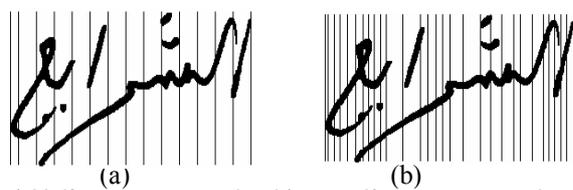


Figure 2. Word segmentation into frames.

The parameter vector of each frame is a combination of 33 statistical and structural features. The first features set is formed of 16 statistical (described in [5]) features that represent distribution features and are computed after dividing each frame into cells with a cell height fixed to 4 pixels. These features are based on foreground (black) pixel densities. Furthermore, some of them are baseline dependent. The next features set is composed of 17 structural features whose 8 concavity features that reflect

local concavity and stroke direction, and also some of these features are baseline dependent [5]. The preceded features are computed from the bitmap representation of word. The 9 other structural features are computed from the thinned image and correspond to 1) feature points (i.e., branch, crossing and end points); 2) inflection point (i.e., a change in curvature); 3) cusp points (i.e., a sharp change in direction, that is, two segments form an acute angle); 4) diacritic points (i.e., black pixel having 0 foreground neighbor) with their location (above or below the base line); 5) Loops with how they are included (partially or completely) in the frame.

6. Word Model Training and Classification

Word model training is carried out so that to build for each word in the lexicon a SCHMM with explicit state duration. This task performs characters model training and the word model is built up by character models concatenation. This makes the system flexible with respect to a change of lexicon because any word is a characters string. In this way, it is sufficient to have in the training set samples of characters composing the words to be modeled rather than samples of the words themselves. Furthermore, the numbers of parameters is kept lower because the word models share the parameters belonging to the same characters. This can improve the training quality given the same amount of training data. In our case, we have not at one's disposal letter samples, however we have word samples, as result we do not apply directly the training algorithm to letter models, but to their concatenations corresponding to the words in the training set. This is called *embedded training* and has two important advantages: the first one is that the characters are modeled when being part of a word (that is the actual shapes of the characters in the cursive handwriting), the second one is that, it is not necessary to segment the words into characters to perform the training.

In our word modeling based on SCHMMs with explicit state duration, the state meaning is associated to a logical notion that is the sub-letter. Each character shape is modeled by SCHMM having 4 states and 25 tied mixtures per state. Also, characters with additional marks (Hamza, Chedda, etc.) and ligatures are labeled and modeled separately. Subsequently, we have up to 160 different SCHMMs related to 28 Arabic basic characters. The recognition lexicon is structured as a tree, this allows efficient sharing of the character HMM-models between the words, and hence reduces the storage space.

The SCHMM topology is right to left with solely transitions to the next state and one allowing for skipping of a single state. The state self-transitions are implicitly modeled by the explicit state duration.

Training and classification are basically done by the aforementioned modified version of the Viterbi algorithm. In the training stage a segmental k-means algorithm [12] is performed. In each iteration only the state-vector assignments resulting from best path obtained from applying the Viterbi algorithm are used to re-estimate model parameters. Moreover, we use the formulas (4) and (8) to parameters re-adjustment of state

duration probability distributions.

7. Experimental Results and Discussion

To test our system, we have carried out several experimentations on the IFN/ENIT [10] benchmark database. The database consists of 26459 Arabic words written by 411 different writers, related to a lexicon of 964 Tunisian town/village names. Four distinct sets (a, b, c, d) are predefined in the database. As it is recommend in [10], 3 sets were used for training and one set for testing. Several tests by altering the following parameters: 1) the distribution of the explicit state duration; 2) the segmentation procedure into frames, were performed.

The Table 1 summarizes the mean results reached following a cross-validation scheme (i.e. we use each time 3 data sets for training and one data set for testing the system). The standard deviations of the obtained results are 1.51 (in top 1), 1.32 (in top 2) and 1.72 (in top 10).

State Duration Distribution	Segmentation Type	Top 1	Top 2	Top 10
Standard SCHMMs	Uniform	81.02	84.14	91.74
	No-Uniform	83.79	85.16	92.12
SCHMMs & Poisson	Uniform	82.53	85.11	93.09
	No-Uniform	84.89	86.17	94.56
SCHMMs & Gauss	Uniform	84.47	87.04	94.88
	No-Uniform	86.10	88.90	95.36
SCHMMs & Gamma	Uniform	87.01	89.78	95.15
	No-Uniform	89.79	92.25	96.78

Table 1. Mean recognition results using each time 3 data sets for training and one data set for testing.

The above results show that SCHMMs with explicit state duration are more efficient for modeling Arabic handwriting compared to standard SCHMMs. The average performance gain is 10.82% in top 1 with Gamma distribution (the best recognition rate in top 1 is 90.89% when using the data sets (a,b,d) for training and the data set (c) for testing). The preeminence of Gamma distribution for state duration can be attributed to its statistical proprieties and to the appropriateness of the data used for estimating its parameters. The discrete Poisson distribution results are less accurate than those of Gauss and Gamma. This fact can be explained by the insufficiencies of the training data, as these need to well estimate the one parameter of Poisson distribution. On the other hand, the no-uniform segmentation scheme is more suitable than the uniform one, because the no-uniform segmentation almost gives rise to a frame whose shape represents a complete character or a sub-character. By contrast, the uniform segmentation can always produce a frame representing a partial combination of 2 characters.

8. Conclusion

In this paper we have proposed a segmentation-free method for Arabic handwritten words recognition using SCHMMs with different explicit distributions for the state

duration, and combining statistical and structural features extracted according two segmentation (uniform and non-uniform) schemes into frames. The results obtained are very promising and have shown that the explicit state duration modeling within SCHMMs framework can improve significantly the recognition rate. Moreover, continues distributions (i.e. Gamma and Gauss) of the state duration are more suitable than discrete ones (i.e. Poisson) for Arabic handwriting modeling, and the no-uniform segmentation scheme is more recommended. The main drawback of SCHMMs is imperfect observation probability estimation. Hence, our foreseen perspective to surpass this problem is searching an appropriated integration method of artificial neural networks (ANNs) to SCHMMs with explicit state duration, for best estimation of observation probability. Also, one can improve the system performance by combining several classifiers (SCHMMs, MLP, SVMs, etc.).

References

- [1] A. Amin, H. Al-Sadoun, and S. Fischer, "Hand-Printed Arabic character recognition system using an artificial network", *Pattern Recognition*, Vol 29, N° 4, pp: 663-675, 1996.
- [2] A. Amin, "Off-Line Arabic character recognition: the state of the art", *Pattern Recognition*, Vol. 31, N°. 5, pp: 517-530, 1998.
- [3] E.N Ben Amara., and F. Bouslama, "Classification of Arabic script using multiple sources of information: State of the art and perspectives". *IJDAR*, Vol 5, N° 1, pp: 195-212, November 2003.
- [4] M. Blumenstein, C.K. Cheng, X.Y. Liu, "New Preprocessing techniques for handwritten word recognition". *Proceeding of the 2nd IASTED Conference on Visualization, Imaging and Image Processing*, pp: 480-484, 2002.
- [5] R. El-Hajj, L. Likforman-Sulem; and C. Mokbel, "Arabic handwriting recognition using baseline dependent features and hidden Markov modeling". *ICDAR (Proceedings of the Eighth International Conference on Document Analysis and Recognition)*, pp: 893-897, 2005.
- [6] F. Farooq, V. Govindaraju, and M. Perrone, "Pre-processing methods for handwritten Arabic documents". *ICDAR (Proceedings of the Eighth International Conference on Document Analysis and Recognition)*, pp: 267-271, 2005.
- [7] X. D. Huang, Y. Ariki, and M. A. Jack, "Hidden Markov models for speech recognition", *Edinburgh Universal Press*, 1990.
- [8] S.E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition". *Computer, Speech & Language*, Vol 1, N° 1, pp: 29-45, 1986.
- [9] T. Pavlidis, "Algorithms for Graphic and Image Processing", *Rockville, MD : Computer science press*, pp : 195-200, 1982.
- [10] M. Pechwitz, S.S. Maddouri, V. Maergner, N. Ellouze, and H. Amiri, "IFN/ENIT - database of handwritten Arabic words". In *Proc. of CIFED 2002*, pages 129-136, Hammamet, Tunisia, October 21-23 2002.
- [11] M. Pechwitz., and V. Maergner, "HMM Based Approach for Handwritten Arabic Word Recognition Using the IFN/ENIT- Database", *ICDAR (Proceedings of the Seventh International Conference on Document Analysis and Recognition)*, pp: 890-894, 2003.

- [12] L.R. Rabiner, "A tutorial on hidden Markov models and select applications in speech recognition", Proceeding of IEEE, Vol 77, N° 2, pp: 257-286, February 1989.
- [13] M. J. Russel, and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov models for speech recognition", Proceeding of ICASSP (International Conference on Acoustic, Speech and Signal Processing), pp: 5-8, 1985.
- [14] A. Vinciarelli, S. Bengio, and H. Bunke, "Offline recognition of unconstrained handwritten texts using HMM's and statistical language models". IEEE Transactions on PAMI, Vol 26, N° 6, pp: 709-720, 2004.