J Inf Process Syst, Vol.10, No.4, pp.503~522, December 2014
http://dx.doi.org/10.3745/JIPS.02.0006

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

Graphemes Segmentation for Arabic Online Handwriting Modeling

Houcine Boubaker*, Najiba Tagougui*, Haikal El Abed**, Monji Kherallah*, and Adel M. Alimi*

Abstract—In the cursive handwriting recognition process, script trajectory segmentation and modeling represent an important task for large or open lexicon context that becomes more complicated in multi-writer applications. In this paper, we will present a developed system of Arabic online handwriting modeling based on graphemes segmentation and the extraction of its geometric features. The main contribution consists of adapting the Fourier descriptors to model the open trajectory of the segmented graphemes. To segment the trajectory of the handwriting, the system proceeds by first detecting its baseline by checking combined geometric and logic conditions. Then, the detected baseline is used as a topologic reference for the extraction of particular points that delimit the graphemes' trajectories. Each segmented grapheme is then represented by a set of relevant geometric features that include the vector of the Fourier descriptors for trajectory shape modeling, normalized metric parameters that model the grapheme dimensions, its position in respect to the baseline, and codes for the description of its associated diacritics.

Keywords—Baseline Detection, Diacritic Features, Fourier Descriptors, Geometric Parameters, Grapheme Segmentation, Online Arabic Handwriting Modeling

1. INTRODUCTION

A major goal of pattern recognition research is to recreate human perception capabilities in artificial systems. Handwriting is an essential human skill, since it is one of the most familiar forms of communication. This is why a pen-based interface combined with automatic handwriting recognition will facilitate the use of mobile devices. It will do so by offering a very easy and natural data entryway [1]. This also explains the increase of interest in handwriting modeling and recognition over the last four decades.

Our paper deals with an integration process for the graphemes segmentation and features extraction of Arabic handwriting acquired online for script modeling and recognition. The baseline, on which the trajectories of aligned and joined characters are located, constitutes a useful segmentation marker for the acquired cursive script [2-4]. To detect it we decomposed the script path in groups of nearly aligned points according to the direction of their tangents to the

※ This work supported by grants from the General Direction of Scientific Research and Technological Renovation (DGRST), Tunisia, under the ARUB program 01/UR/11/02.
Manuscript received July 31, 2013; accepted November 17, 2013; onlinefirst October 7, 2014.
Corresponding Author: Houcine Boubaker (houcine-boubaker@ieee.org)
* Research Groups in Intelligent Machines (REGIM), National School of Engineers ENIS, University of Sfax, BP 1173, Sfax 3038, Tunisia. (Houcine-boubaker@ieee.org, tag_najiba@yahoo.fr, monji.kherallah@enis.rnu.tn, adel.alimi@ieee.org)
** Institute for Communications Technology (IfN), Technische Universität, D-38106 Braunschweig, Germany. (elabed@tu-bs.de)

trajectory. Then, the extracted sets of points were tested to verify the topological conditions to evaluate their relevance as a support for the baseline.

The segmentation module receives the preprocessed handwriting trajectory and the data defining the baseline direction of each pseudo-word as input. It extracts two types of particular points, which are the bottom of the valleys close to the baseline and the vertical trajectory turn back summit, for the fragmentation of handwriting that is in a basic shape called graphemes. The shapes of the segmented graphemes and their position in respect to the baseline are then modeled by extracting the relevant features vector. From parametric or structural techniques, the choice was for a mixed model: parametric features for graphemes structural entities representation, to ensure that the model can be adapted for fuzzy recognition approach with either limited or open lexicon context. Fourier descriptors are one of the most accurate tools for the parametric modeling of a closed path, which can be represented by a $2 \cdot \pi$ periodic signature function [5]. They are successfully used to model the closed contours of the connected components area in the treatment process of digital images. However, taking advantage of their approximation aptitude in grapheme modeling requires the transforming of a non-periodic signature, corresponding to the open trajectory of a segmented grapheme, to a periodic function. On the other hand, we also needed to conduct a study to choose the appropriate number of harmonics k to consider for the approximation of the original signal by a Fourier series, which will allow us to define the number of Fourier descriptors that need to be inserted into a parametric model.

The features vectors is also enhanced by including other normalized geometric parameters that measure the dimension of grapheme and the positions of its trajectory extreme points in respect to the baseline and features representing its associated diacritics.

In the evaluation phase, we applied the system to the Arabic handwriting database (ADAB) [6] of the names of Tunisian towns online. We did so by using a classifier module based on hidden Markov models (HMMs) that we implemented via the HMM Toolkit.

This paper is organized as follows: in Section 2 we present the state-of-the-art. In Section 3 we describe the preprocessing and the used baseline detection algorithm. In Section 4 we present the graphemes segmentation method. Section 5 illustrates the developed modeling approach and the extracted grapheme features vector. In Section 6 we conclude by presenting the experimental tests results and the perspectives of application and improvement of the developed modules.

2. RELATED WORK

The modeling and recognition approaches that perform an explicit segmentation in characters or graphemes for cursive handwriting are defined as analytical approaches [7,8]. Their advantage is that they are able to work on an extended or infinite lexicon.

Prior to any modeling or recognition step, the acquired data is generally preprocessed to reduce noise, to normalize the various aspects of the trace, and to segment the signal into meaningful units. For Arabic cursive handwriting, baseline detection is a step that is principally used for the detection of delay strokes [9,10] or character segmentation [4,8] and features extraction [11].

The geometric approaches of baseline detection, such as histogram projection [3,7], the Hough transform [12], or the entropy method [7], apply geometric transformation onto blocks of script that are large enough. This is done so as to detect the direction according to which the handwriting entities are aligned and appear mostly compact [7]. Conversely, the logic approaches

of baseline detection are preceded by a topological analysis of the handwritten script to discern or to select the relevant points (trajectory vertical extremes [13]) or stroke shape (valley [13], loops [4]) of the trajectory that supports the searched baseline.

Segmentation refers to the different operations that must be performed to get the basic entities of the handwriting that the recognition algorithm will have to process. It generally works on two levels. The first level deals with the entire text and focuses on line detection [14,15] or more precisely—word segmentation. It proceeds by the detection of spatial zones or temporal order or both. At the second level, the methodology focuses on the segmentation of the input data into individual characters or even into sub-character units, such as strokes or graphemes. This operation is one the most challenging aspects, particularly for the recognition of cursive script.

Many segmentation techniques have been developed for the modeling and recognition of Arabic online handwriting, as discussed by Abuzaraida et al. [16].

Izadi et al. [17] decomposed the signal into convex/concave segments that represent elementary shapes. In order to avoid finding segments of very short lengths, a threshold is applied to the length of the segment curve. This represents the sum of the lengths of the piecewise linear segments that construct the curve. Daifallah et al. [18] presented an algorithm that works on strokes and segments them into letters in the following four stages: arbitrary segmentation, segmentation enhancement, connecting consecutive joints, and locating segmentation points. Eraqi and Abdelazeem [9] segmented the pseudo-words in graphemes based on the detection of significant points. Their algorithm depends on the local writing direction and is independent of the baseline so as to be less sensitive to baseline detection errors. Sternby et al. [19] invoked a segmentation technique based on the principle of Frame Deformation Energy, where each stroke is subdivided into segments based on the orthogonal direction of the writing direction using a set of heuristic rules.

Elanwar et al. [10] proposed a segmentation-recognition procedure by using a dynamic programming algorithm to find a globally optimal set of cuts through the input test string (feature vector), which minimizes the defined cost function.

In the handwriting recognition field, many different techniques have been used. In [20], the authors presented a connectionist approach. This uses a bidirectional single recurrent neural network with long short-term memory architecture that uses a function, known as a Connectionist Temporal Classification, which uses the network to label the entire input sequence all at once in a way to directly train the network to label unsegmented sequence data. This system was tested on the IAM online handwriting database (IAM-OnDB), and achieved a word recognition rate of 74.0%. This technique is also used by Graves et al. [21].

In order to have better recognition accuracy, Liwicki and Bunke [22] proposed to not only use one recognition classifier, but to combine several individual recognition systems based on HMMs and bidirectional long short-term memory (BLSTM) networks that use various feature sets based on online and offline approaches. In [23], the handwritten data is recognized by using continuous HMMs (each HMM models one character) after the identification of script lines. The character level accuracy is 63.3% and the word level accuracy is 64.8%.

For the recognition of online Arabic handwriting most studies have focused more on classification mechanisms than on other recognition aspects, like signal modeling. We noticed the great interest in the development of online handwriting recognition systems in [24-28]. A variety of representations or signal modeling of isolated characters or that are assumed to be the result from a reliable segmentation stage have been used, such as a decomposition into

characteristic strokes [29]; global shape descriptors, such as Fourier coefficients [30]; and local geometric descriptors, such as tangents [31]. For other scripts that are online, the coordinates of the input signal [24,32,33], have also been used to extract time-dependent representation features, such as curvilinear and angular velocities [26,34,35]. All of these representations of form or pattern describe plausibly handwritten Arabic cursive script.

Respect to the state-of-the-art, the online Arabic handwriting modeling system that we are presenting addresses a mixed approach that represents the handwriting structural entities (graphemes) with a parametric features model that combines Fourier descriptors; geometric parameters, which measure the grapheme dimensions; and location, in respect to the baseline and codes that represent the grapheme associated diacritics.

3. PRE-PROCESSING AND BASELINE DETECTION

3.1 Size Normalization and Trajectory Filtering

We have applied the developed grapheme modeling system to online Arabic handwriting. A digital tablet, or a similar type of device, captures the handwriting trajectory. It may represent short messages, replies to an electronic form, notes on electronic agenda, lessons, e-mail, etc. The digital sampled trajectory of the pen is represented as a function of time of its points coordinates. The preprocessing stage aims to normalize the handwriting dimensions (l_H, l_V) and to eliminate the noise. First, the vertical dimension l_V of the handwritten line sentence is adjusted to a fixed value L_{V_norm} in order to obtain a normalized size script (L_H, L_V) (see Fig. 1).

$$L_V = L_{V\_norm} \text{ and } L_H = l_H \cdot \frac{L_{V\_nor}}{l_V} \tag{1}$$

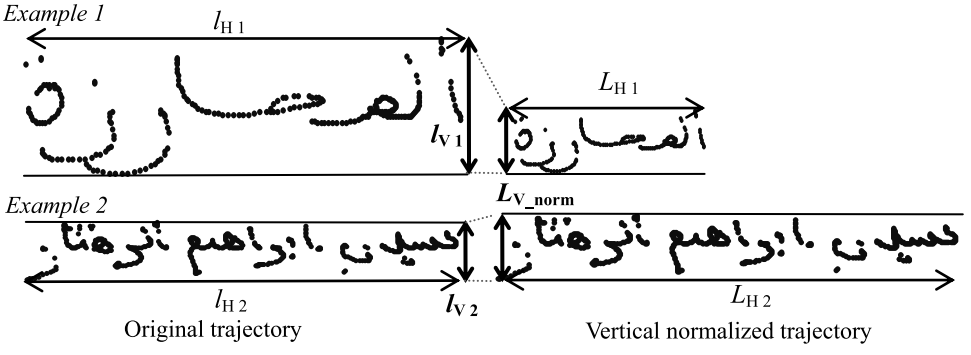


Fig. 1. Vertical dimension normalization.

Then we applied a second Chebyshev low-pass filter with a cutoff frequency of f_{cut}=12 Hz to the normalized trajectory in order to eliminate the noise introduced by temporal and spatial sampling [36].

3.2 Baseline Detection Algorithm

The baseline constitutes a topological marker for the segmentation of cursive handwriting in basic shape representing characters or graphemes [2,3,12,13]. The algorithm that we used combines Geometric method with Logic procedure for handwriting Baseline Detection [37]. This is concretized by the following two steps: first, it detects a starting set $\{M_i\}_{Str}$ of trajectory points M_i having a nearly horizontal tangent direction that will be decomposed in q groups of nearly aligned points $G_j \quad j=1,\ldots,q$ (see Fig. 2) by affecting each point of M_k to the already constituted group G_j where it minimizes the mean of the two distances listed below.

- $D_{M_k G_j}$: the distance between the point candidate M_k and the regression line representing the group of points G_j.
- $D_{C_j T_k}$: the distance between the centroïde C_j of the group G_j and the tangent of the trajectory on the point M_k.

In the case where the two measured distances exceed a maximal limit value, the point candidate M_k is assigned to initialize a new group of points incrementing consequently the number q.

In the second step, the three most extended groups of points (in terms of the number of elements) extracted from the first step are then tested to measure their conformity to the topological conditions and rules that are specific to the Arabic handwriting baseline (listed below) and to evaluate their relevance to be recognized as such.

- Low angle of intersection between the upward trajectory and baseline.
- Reduction of the average angle of the absolute curvature of the graphemes segmented in respect to the assumed baseline.
- Concentration of the contact points between handwriting strokes and the selected baseline on the trajectory middle part more than toward its endpoints.

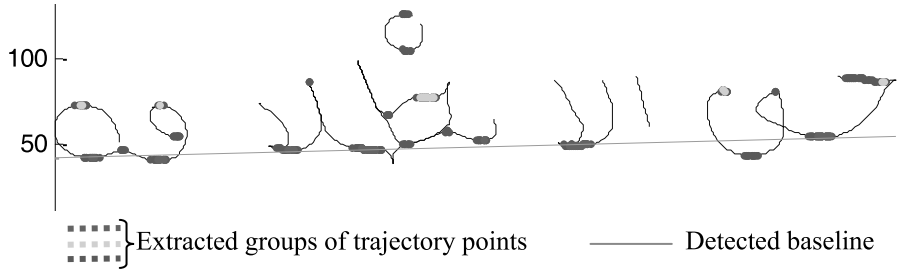


Fig. 2. Example of the extraction of groups of points and baseline detection result.

Tests carried out on online handwritten script from the ADAB database [6] give about 97.4% of goods baseline detection results (see Fig. 2).

4. Graphemes Segmentation of Handwriting

The operation of handwriting segmentation is one of most challenging aspects, particularly for the recognition of cursive script [1]. The detection of the baseline by the above presented algorithm permits to define the virtual line, on which the cursive handwriting characters are

arranged and/or joined. The inter-graphemes ligature valleys are localized in the horizontal median zone shared by all concatenated or isolated graphemes [38]. The estimation of the thickness of this zone permits to define a neighborhood around the baseline in what is estimated to be the presence of the ligature valleys.

4.1 Estimation of the Width of the Median Zone

For Arabic writing, the middle zone represents the most shared and thickest horizontal level of a line of script [38]. Its vertical width can be identified by analyzing the horizontal projection histogram [7,39,40] (see Fig. 3). Given that in our application the detected baseline may not be straight, the projection of the part of each continuous handwriting stroke (COHS)—handwriting trajectory limited between pen-down and pen-up moments— located above the baseline, is made according to its corresponding local baseline direction. This process conducts to the computation of an elementary horizontal profile $H_i(n)$ for each i^{th} COHS. In order to quantify the positions of the projected points, we divided the distance on the profile vertical axis that is limited between the upper limit line and the baseline in *N_int* intervals. The histogram of the horizontal projection of the all text lines composed of j COHS is then obtained by the sum of their j elementary horizontal profiles:

$$H(n) = \sum_{i=1}^{j} H_i(n) \quad \text{for} \quad n = 1, ..., N\_int \tag{2}$$

The estimation of the median zone upper level (median line) consists in looking for the vertical level, which maximizes the derivative of the horizontal profile (see Fig. 7):

$$Diff(n) = H(n) - H(n+1) \quad \text{for} \quad n = 1, ..., N\_int \tag{3}$$

Thus, the thickness of the median zone (h_{ZM}) is obtained by measuring the distance between the baseline and the median line.

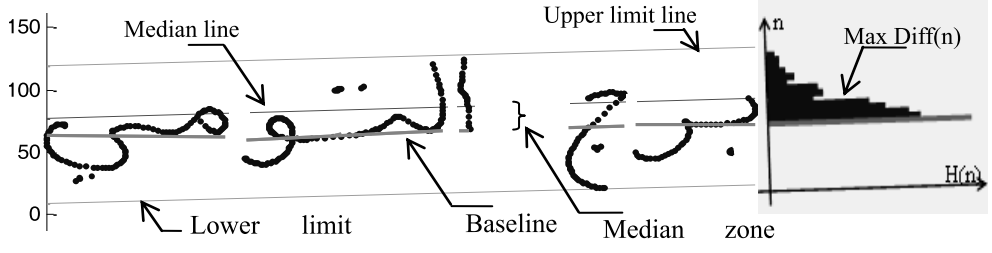


Fig. 3. Median zone detection by horizontal projection of the trajectory strokes handwritten above the baseline.

4.1 Detection of the Graphemes' Limits

The term 'graphemes' defines the set of basic graphic shapes that cursive text is composed of. One grapheme can represent a whole character or a section of its tracing. For example, several

Arabic characters such as 'س' , 'ب' , 'ت' include one or several graphemes named 'nabra' 'د'. The segmentation of the Arabic pseudo–words in graphemes is based on the detection of two types of topologically significant points (M_{PP}) (see Fig. 4).

- The bottom of the ligature valleys: this is the point of a trajectory segment moving from right to left that verifies the most local and closest position to the baseline with trajectory tangents that are parallel to its direction.
- The angular point: this is the top point that represents the extremum of a vertical shaft trajectory turning back.

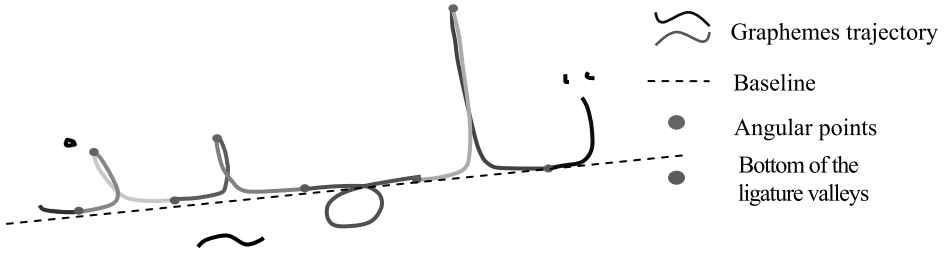


Fig. 4. The topologically significant points.

Each M_{tm} point corresponding to a local minimum of the absolute tangent deviation angle in respect to the baseline: $\Delta\alpha(i) = \left| \alpha_{tg\,M_i} - \alpha_{baseline} \right|$ (where $\alpha_{baseline}$ is the baseline slant angle), is considered to be a particular point candidate M_{PP}. It will be kept as a particular point M_{PP} if it is enough close to the baseline with a tangent to the trajectory that is almost parallel to it (bottoms of ligature valleys) or if it corresponds to a sharp deviation of the trajectory with an almost vertical median direction (angular points) (see Fig. 5). In our experiment we retained the particular points that verified the following empiric and topological conditions:

$$\text{bottom of the ligature valleys} \begin{cases} R_{\Delta y} = \dfrac{\Delta y}{h_{ZM}} < R_{\Delta y\,max} = \dfrac{1}{2} \\ \Delta\alpha < \Delta\alpha_{max} = \dfrac{\pi}{6} \end{cases} \tag{4}$$

$$\text{angular points} \begin{cases} \Delta\theta > \Delta\theta_{min} = \dfrac{\pi}{2} \\ Dev\theta_{med} = \left| \theta_{med} - \dfrac{\pi}{2} \right| < Dev\theta_{max} = \dfrac{\pi}{5} \end{cases} \tag{5}$$

where $R_{\Delta y}$, is the normalized ratio that represents the position of the point M_{tm} in respect to the baseline. Δy is the distance between M_{tm} and the baseline. h_{ZM} is the width of the median zone. $\Delta\alpha$ is the deviation angle of the trajectory tangent on the M_{tm} point in respect to the baseline. For the topological conditions (5) that have been reserved for the detection of the angular points; $\Delta\theta$ is the deviation angle between the direction of the tangents in the respective trajectory neighborhoods located before and after the M_{tm} point. $Dev\theta_{med}$ is the deviation angle connected to the vertical of their median direction (see Fig. 5).

The values of the thresholds $R_{\Delta y\,max}$, $\Delta\alpha_{max}$, $\Delta\theta_{min}$ and $Dev\theta_{max}$ are retained as statistical analyses results of experimental tests presented in Subsection 6.1.1.

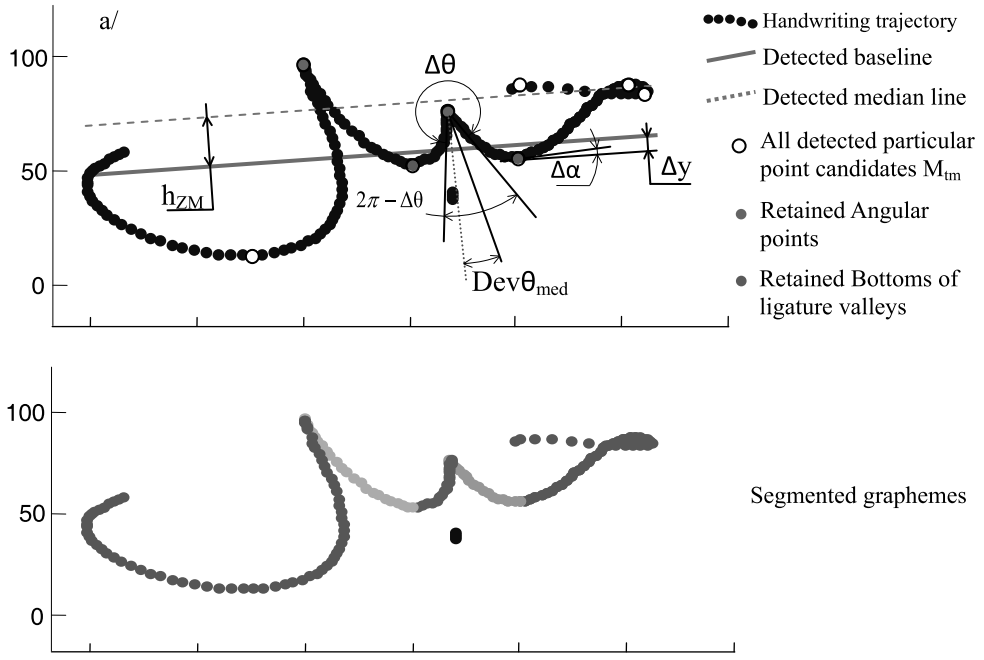


Fig. 5. Topological characteristics examined for (a) the detection of particular points and (b) grapheme segmentation results.

5. GRAPHEME MODELING

The objective of this section consists in extracting relevant parametric features that characterize the shape, dimensions and position of each segmented grapheme from the handwritten script. The used features set includes:

- Fourier descriptors for grapheme trajectory shape modeling.
- Geometric parameters for grapheme location and trajectory endpoints and maximum curvature point marking.
- Representation of diacritics.

5.1 Fourier Descriptors for the Modeling of Grapheme Trajectory Shapes

Fourier descriptors represent one of the most accurate tools for closed contour modeling [5]. To benefit from their powerful capacity of periodic function approximation in segmented graphemes modeling, we must transform the signatures corresponding to the graphemes open trajectories into periodic functions. Let M_1 and M_n respectively the start and the end points of the grapheme trajectory. The approach that we adopted to solve the problem of periodicity consists in running through the grapheme path in reverse directions; first from M_1 to M_n and then backtracking from M_n to M_1 (see Fig. 6).

The chosen function that we used as a grapheme trajectory signature $\theta_i = f(\ell_i)$ represents the variation of the inclination angle θ_i of the trajectory tangent at the point M_i in accordance to its corresponding curvilinear abscissa ℓ_i:

$$\ell_i = \sum_{j=1}^{i} dL_j \qquad \text{for} \qquad i = 1,...,2n \tag{6}$$

where dL_i is the elementary curvilinear distance between the current point M_i and its previous:

$$\begin{aligned} dL_1 &= \ell_1 = 0 \\ dL_i &= \left\| M_i\, M_{i-1} \right\| \qquad \text{if} \qquad 1 < i \le n \\ dL_i &= \left\| M_{2n-i+2}\, M_{2n-i+1} \right\| \qquad \text{if} \qquad n+1 \le i \le 2n \end{aligned} \tag{7}$$

The Fourier series can approximate the defined grapheme signature, since it is a periodic (and symmetrical) function that verifies:

$f(\ell_1) = f(\ell_{2n}) = \theta_1 =$ trajectory tangent inclination angle at the point M_1,

$f(\ell_i) = f(\ell_{2n-i+1}) = \theta_i =$ trajectory tangent inclination angle at the point M_i for i =1, …,n (see Fig. 6).

We then calculated the Fourier descriptors parameters, which constitute the coefficients a_0, a_j and b_j for j=1,…, k, of the Fourier series that approximate, at the k^{th} harmonic, the signature function $\theta_i = f(\ell_i)$

$$a_0 = \frac{1}{2 \cdot \pi} \cdot \sum_{i=1}^{2n} \theta_i \cdot \, dL_i \tag{8}$$

$$\begin{cases} a_j = \dfrac{1}{\pi} \cdot \displaystyle\sum_{i=1}^{2n} \theta_i \cdot \cos\left(j \cdot \frac{2 \cdot \pi \cdot \ell_i}{\ell_{2n}} \right) \cdot dL_i \\ b_j = \dfrac{1}{\pi} \cdot \displaystyle\sum_{i=1}^{2n} \theta_i \cdot \sin\left(j \cdot \frac{2 \cdot \pi \cdot \ell_i}{\ell_{2n}} \right) \cdot dL_i \end{cases} \qquad j=1,\ldots,k \tag{9}$$

For the reconstruction of both the grapheme signature and trajectory, we used the following approximation function that is comprised of the Fourier series:

$$\theta_i = f(\ell_i) \;\approx\; a_0 + \sum_{j=1}^{k} \left[a_j \cdot \cos\left(j \cdot \frac{2 \cdot \pi \cdot \ell_i}{\ell_{2n}} \right) + b_j \cdot \sin\left(j \cdot \frac{2 \cdot \pi \cdot \ell_i}{\ell_{2n}} \right) \right] \tag{10}$$

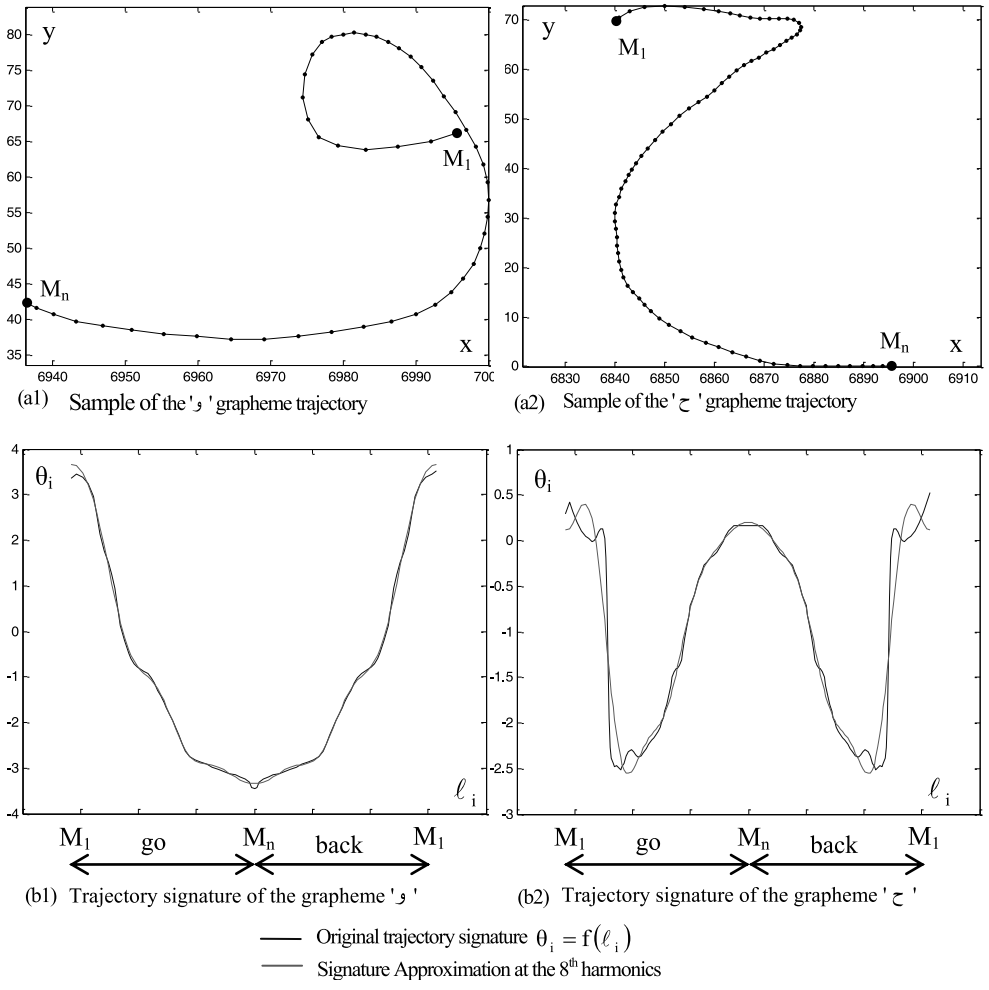


Fig. 6. Graphemes trajectories (a1, a2) and the approximation of corresponding signature functions at the 8th harmonic (b1, b2).

The choice of the appropriate number of harmonics k will be discussed in the tests and results in Section 6.

5.2 Geometric Parameters of Grapheme Location and Trajectory Marking

The Arabic letters or graphemes can be partially characterized by their measurements (vertical and horizontal dimensions L_V, L_H) and the location of their trajectories in respect to the baseline. For example, the graphemes '١' and 'ب' are quite distinct considering that only the dimensions the smallest quadrilateral rectangles (called a bounding box) [13] can surround all the points $M_i(x_i, y_i)$ composing each grapheme and for which an edge is parallel to the baseline (see Fig. 7). On the other hand, the levels of the bounding box's vertical edges respect to the baseline, allow to distinguish the graphemes that are written in over of the baseline 'د , ف , …' from those that descend underneath the baseline 'ر , ز , ن , …' and from the character diacritics.

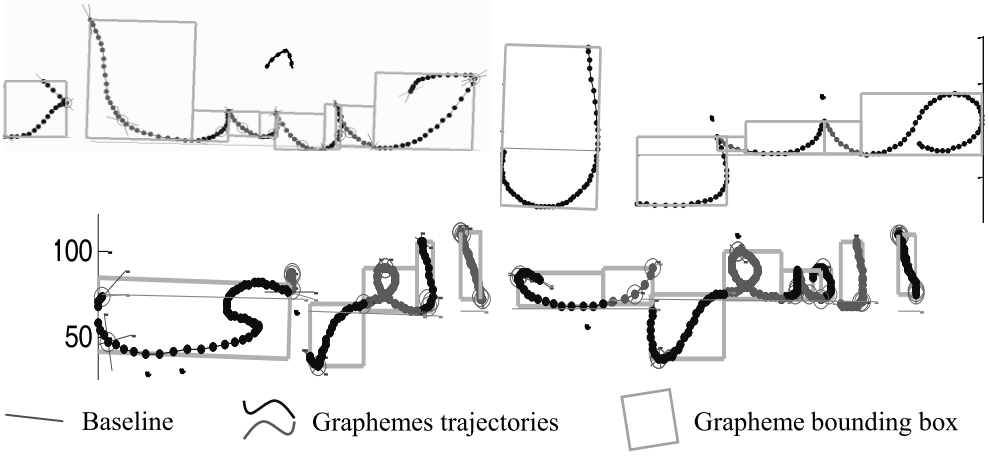


Fig. 7. A grapheme's bounding box delimitation.

The grapheme trajectory is then marked by three extracted marking points:

- The grapheme trajectory starting point M_1.
- The grapheme trajectory end point M_n.
- The point $M_{min} \in]M_1, M_n[$ corresponding to the absolute minimum of trajectory curvature radius $R_{C\_i} = \frac{\Delta \ell_i}{\Delta \theta_i}$ (see Fig. 8).

The positions of the three marking points in the grapheme bounding box give an overview on the shape of the grapheme trajectory. These positions are defined in respect to the left lower summit $S_{bb}(x_S, y_S)$ of the bounding box in the horizontal and vertical directions by the ratios R_H and R_V.

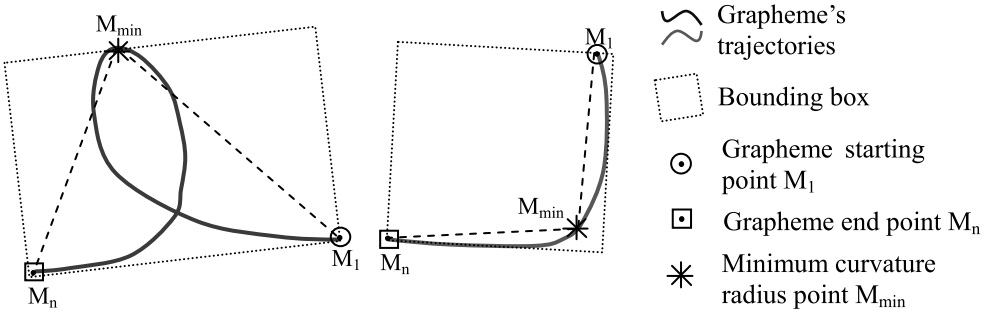


Fig. 8. Position of the trajectory marking points on the grapheme's bounding box.

To introduce more precision to the trajectory model, we also determined the angles of the trajectory tangent slant θ_1, θ_i, and θ_n, on the respective three marking points of M_1, M_{min}, and M_n and the algebraic values of the curvature angles α_{Ca_1}, α_{Ca_2} of the grapheme trajectory stroke draw before and after the M_{min} extremum point.

5.3 The Representation of Diacritics

The strokes of the diacritics are first filtered from the main handwritten script by examining the measurements and the positions of their trajectories in respect to the baseline. The detected diacritics are then analyzed and classified according to their sizes and shapes that are modeled by the Fourier descriptors as single dot, two merged dots, three merged dots, or 'shadda' 'ّ' using a k-nearest neighbors algorithm. The resultant numbers of the merged or discrete upper and lower diacritic dots associated to each segmented grapheme and its rate of association of the diacritic 'shadda' are inserted into the grapheme features vector.

6. EXPERIMENTS AND RESULTS

In the evaluation phase, we applied the system to the ADAB database [6,41,42], which includes the names of 937 Tunisian towns entered online. Details of different sets are presented in the Table 1.

Table 1. Statistics on the different sets that the ADAB database is composed of

Set	File	Word	Character	Writer
1	5,037	7,670	40,500	56
2	5,090	7,891	41,515	37
3	5,031	7,730	40,544	39
4	4,417	6,786	35,832	25
5	1,000	1,551	8,189	6
6	1,000	1,536	8,110	3
Sum	21,575	33,164	174,690	166

6.1 The Estimation of Variables and Stability Analysis

6.1.1 Choice of a particular point's detection thresholds

Since the detection of the particular segmentation points, the bottoms of the ligature valleys and the angular points (extremum of a vertical shaft trajectory turning back) depends on the precision of the respective thresholds $R_{\Delta y\ max}$, $\Delta\alpha_{max}$, $\Delta\theta_{min}$ and $Dev\,\theta_{max}$ used in the topological conditions expressed by Formulas (4) and (5), the estimation of the values of these thresholds must be based on the statistical results of large experimental tests.

To estimate the value of the $R_{\Delta y\ max}$ threshold we tested the system on a set of handwriting samples, including 6,135 valleys ligature adjoining the baseline, 742 leg valleys or pockets below the baseline, and 186 diacritical valleys or middle zone valleys above the baseline. The rates of ligature valleys correct and false detection are calculated for different values of the $R_{\Delta y\ max}$ threshold going from 0 to 1 by a step of $\frac{1}{8}$ (see Fig. 9). The rate of correct detection reaches its maximum for a $R_{\Delta y\ max}$ value that is close to 0.5 before decreasing by confusing valleys of another level as a ligature valley.

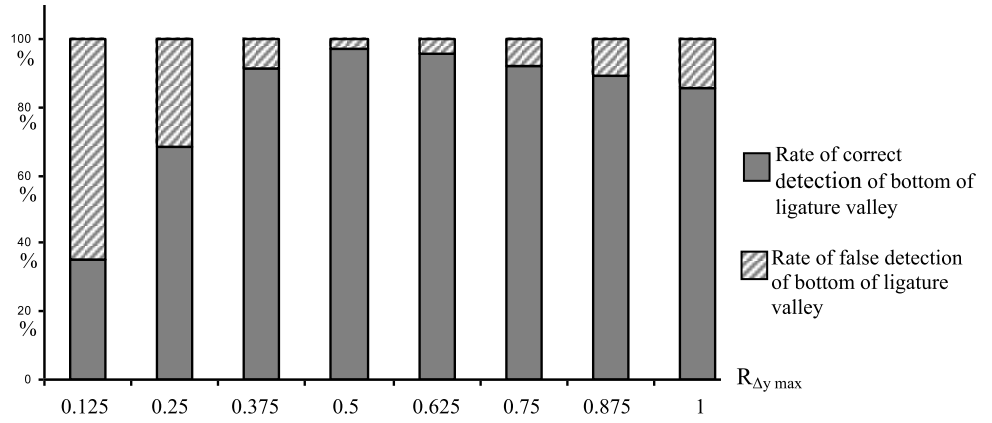


Fig. 9. The correct and false detection rates of ligature valleys according to different values of the $R_{\Delta y\ max}$ thresholds.

The value of the thresholds; $\Delta\alpha_{max} = \frac{\pi}{6}$, $\Delta\theta_{min} = \frac{\pi}{2}$ and $Dev\theta_{max} = \frac{\pi}{5}$ are selected by examining the limits that distinguish the distribution of the segmentation point samples; by the bottoms of ligature valleys and angular points in respect to other trajectory extremum points in the respective maps defined by the pairs of coordinate features $(R_{\Delta y}, \Delta\alpha)$; and by $(\Delta\theta, Dev\,\theta_{med})$ (see Fig. 10).

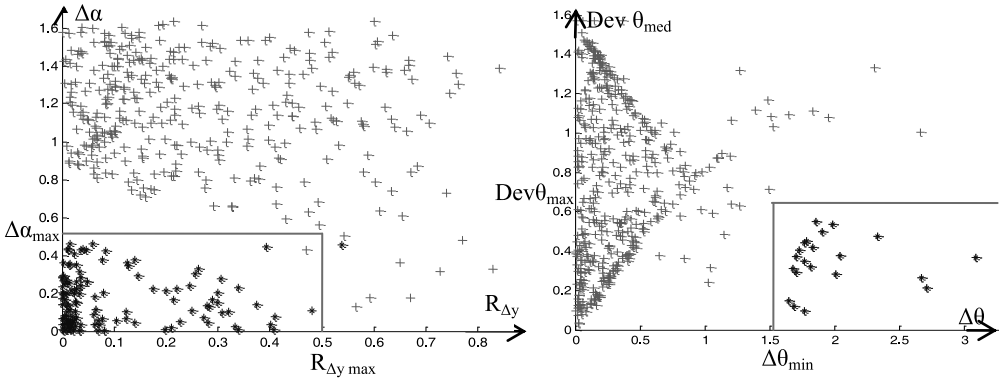


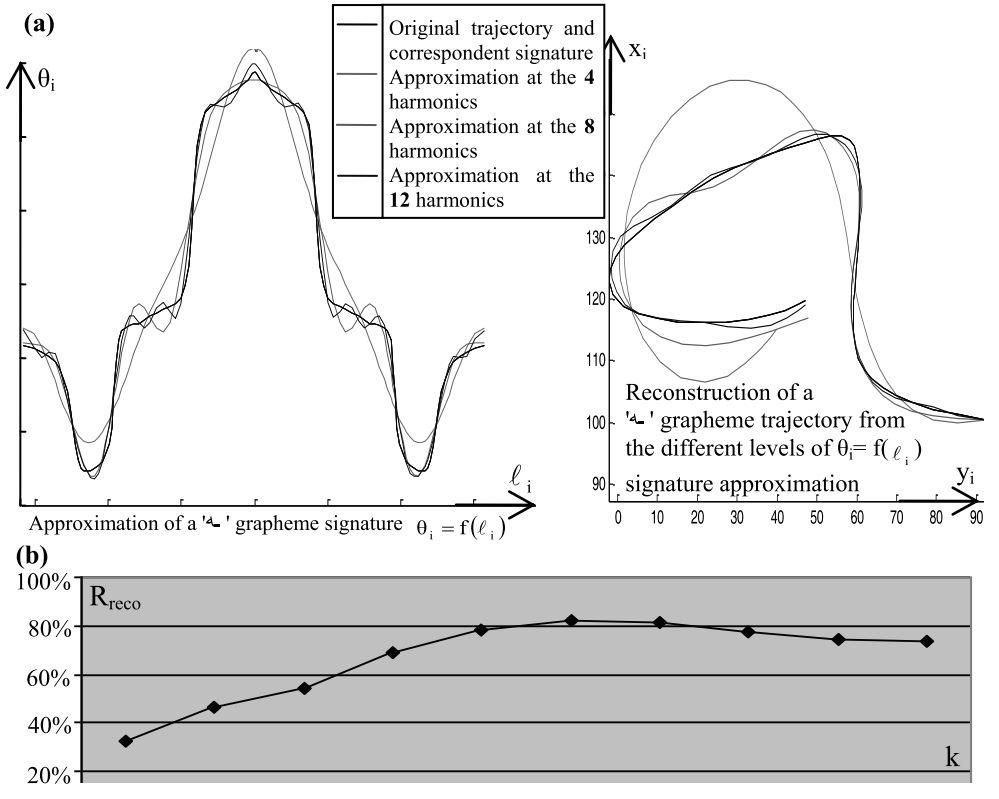
Fig. 10. Maps of trajectory tangent extremum points distribution respect to the pair of feature coordinates $(R_{\Delta y}, \Delta\alpha)$ and $(\Delta\theta, Dev\,\theta_{med})$.

6.1.2. Effect of the number of the Fourier descriptors' harmonics k on performance recognition

In a multi-writer context, the choice of the number of Fourier descriptors' harmonics k is a compromise between precision and generalization. Indeed, when the number of harmonics taken into account increases, the approximation of the grapheme signature $\theta_i = f(\ell_i)$ becomes more

accurate if its resolution in terms of the number of points (M_i, θ_i) i =1,…,*n* is sufficient (the number of points *n* must be greater then 2k). At a given level, this accuracy allows a better distinction between the modeled graphemes. However, when the precision increases further, the obtained parametric model composed by the coefficients a_j and b_j contains data that models the writing style specific to the writer or even the noise at the harmonics of high frequency.

A statistical study has been conducted to determine the most relevant value of k for our application for multi-writer handwriting recognition. It consists in calculating the recognition rate R_{reco} obtained on the same set of 15,138 word samples of the ADAB database written by 130 writers, for different values of k ranging from 3 to 12. It only considers the features vector of the Fourier descriptors a_0, a_j and b_j for j=1,…,k. The obtained results are given by the function curve R_{reco} = f(k) presented in Fig. 11. We denoted that the maximum of the recognition rate curve coincides with the value of k=8. As such, in our application we kept a number of harmonics k=8 (see Fig. 11).



k	3	4	5	6	7	8	9	10	11	12
R_{reco}	32.6%	46.7%	54.4%	69.2%	78.6%	82.3%	81.5%	77.8%	74.1%	73.8%

Fig. 11. Fourier descriptors for grapheme shape modeling. (a) Approximation of the signature $\theta_i = f(\ell_i)$ and the original trajectory of the grapheme ' ' at the 4th, 8th and 12th harmonics, respectively. (b) Variation of the recognition rate, R_{reco}, according to the number of harmonics k.

6.2 Description of Experiments

Our system of online handwriting modeling went through several steps of change and improvement. In each step the system was tested to verify the effects of the introduced changes on the discriminating power of the system.

In its original version [11], the system used a features vector of 21 parameters that was constituted only by the geometric parameters of grapheme location and trajectory marking (see Section 6.2). The trajectory filtering was more suitable to the resolution of the reconstructed skeletons of offline handwriting than to the ripple frequency of online handwriting because the extractor was used as a dual system offline/online [11]. In the second version, we used a Chebyshev low-pass filter (second type) at a cutoff frequency f_{cut} of 12 Hz and a filtering window radius of R = 8. The features vector of graphemes was boosted by the introduction of parameters representing diacritics. The third version of the system was a version of a test for the Fourier descriptors features that we used to select the optimal number of harmonics (see Subsection 6.1). The fourth version used a features vector that combined the geometric parameters of localization and marking with the Fourier descriptors, and the features of diacritics.

All of these different versions of the developed handwriting modeling system were tested under the same conditions in order for us to be able to study and compare the effects of the introduced changes. In fact, we kept the same structure for the classifier to the output of the four modeling system versions that we trained using the first three sets of the ADAB database and tested on its fourth set. The classifier module is a network of interconnected discrete HMMs implemented through the HMM Toolkit as described in [43]. The HMMs that we used were implemented through the left to right discrete topology. The size of the codebook was fixed to 256 codes after several experimental tests. We used the Viterbi algorithm to train the proposed HMM system with the maximum amount of likelihood criterion. The system includes 360 mixtures with 36k Gaussian densities.

6.3 Results and Discussions

The results in terms of the recognition rate obtained by the successive system versions are presented in the Table 2.

Table 2. Results of the successive versions of the system on the different sets of the Arabic handwriting database as given in percentages

System version		Version 1	Version 2	Version 3	Version 4
Training set n°1	Top 1	57.87	86.38	82.33	88.40
	Top 5	72.84	96.43	93.47	98.23
Training set n°2	Top 1	54.26	83.55	80.61	86.27
	Top 5	66.38	94.68	91.53	97.05
Training set n°3	Top 1	53.75	81.26	78.31	87.76
	Top 5	72.31	91.57	90.68	97.83
Test set	Top 1	52.67	77.27	73.56	85.37
	Top 5	63.44	88.35	84.19	96.25

The comparison between the results of versions 1 and 2 shows an improvement in the discriminating power of the system. This is due mainly to the consolidation of the features

vector, which was achieved by introducing the parameters that represent the features of the diacritics, as well as to the calibration of the filtering parameters. The third version of the system that we used as a test to choose the optimal number of harmonics show lower performance than the association of the geometric features and features of the diacritics. The fourth version, which used a parameters vector representing the association between geometric parameters, Fourier descriptors, and diacritic features, improves substantially the recognition results achieving an average rate of 87.46% for the learning sets and 85.37% for the test set.

In Table 3, we present a comparison between the results achieved by the last version of our system and the results of the systems that have participated in the ICDAR 2011 competition.

Table 3. Results of the different systems on the test sets 5 and 6 of the Arabic handwriting database in the ICDAR 2011 competition

System	Set 5			Set 6		
	Top 1	Top 5	Top 10	Top 1	Top5	Top 10
AUC-HMM1	83.13	95.89	96.47	90.40	95.80	96.20
AUC-HMM2	83.33	95.03	95.64	89.90	94.60	95.00
I-H1	62.06	81.71	85.51	66.06	83.70	87.21
I-H2	67.30	83.20	85.82	71.20	87.50	89.20
V-O1	98.89	99.18	98.18	98.45	98.97	98.97
V-O2	98.02	98.13	98.13	98.11	98.55	98.55
Our system	85.37	96.25	98.46	87.62	97.27	98.72

The results of the latest version of our system are slightly lower then those achieved by the first system that competed for online Arabic handwriting recognition in the ICDAR 2011 competition and that was tested on the same ADAB database as our developed system. Conversely, our system is distinguished by its adaptation to applications of large or unlimited lexicons. This is thanks to the explicit segmentation into graphemes that it performs, which allows to conduct an initial level of classification on segmented graphemes. Other qualities may be cited as its less sensitivity to the horizontal variation of ligature elongation of 'madda' or white space [44]. This is thanks to the segmentation strategy inspired from the topological rules for the concatenation of Arabic characters for cursive handwriting.

7. CONCLUSION AND FUTURE WORKS

In this paper, we have presented a system for modeling the Arabic online handwriting based on the segmentation of graphemes. The system consists of the following three modules: baseline detection, grapheme segmentation and features extraction. The first module is characterized by the consideration of geometric and logic features for baseline detection. The second module uses the obtained baseline and the width of the writing median zone to extract particular topological points for the segmentation of the handwriting trajectory in graphemes. The third module extracts a set of parameters combining Fourier descriptors; geometric location parameters; and markings for the modeling of the shape, position, and the associated diacritics of each segmented grapheme. The experimental results show a progressive improvement of the recognition rate with the introduction of new discriminative features. As a continuation to this project, two studies are currently underway. The first one focuses on the optimization of the classification strategy. The second study aims to exploit the presented explicit grapheme segmentation approach to develop

a handwriting recognition system that is applicable on a wide or open lexicon.

REFERENCES

[1] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: a comprehensive Survey," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[2] M. Pechwitz, and V. Märgner, "Baseline Estimation For Arabic Handwritten Words," in *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition IWFHR*, Ontario, Canada, 2002, pp. 479 – 484.

[3] S. Snoussi Maddouri, F. Bouafif Samoud, K. Bouriel, Noureddine Ellouze, "Baseline extraction: comparison of six methods on IFN/ENIT database," in *Proceeding of the International Conference on Frontiers in Handwriting Recognition*, Montréal, Canada, 2008, pp. 571–576.

[4] C. Olivier, H. Miled, K. Romeo, Y. Lecourtier, "Segmentation and Coding of Arabic Handwritten Words," in *Proceeding of the International Conference on Pattern Recognition 13th ICPR*, Vienna, Austria, Oct. 1996, pp. 264–268.

[5] E. Persoon and K. S. Fu, Shape "Discrimination using Fourier descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 3, pp. 388–397, 1986.

[6] M. Kherallah, N. Tagougui, Adel M. Alimi, H. Elabed, V. Märgner, "Online Arabic Handwriting Recognition Competition," in *Proceedings of the 11th International Conference on Document Analysis and Recognition*, Beijing, China, 2011, pp. 1454-1459.

[7] B. Al-Badr and S. A.Mohmond. "Survey and bibliography of Arabic optical text recognition," *Signal Processing*, 1995, pp. 49–77.

[8] S. Garcia-Salicetti, B. Dorizzi, P. Gallinari, Z. Wimmer, G. Toussaint, "A hybrid Neural Predictive Model for On-Line Handwriting Recognition," in *World Multiconference on Systemics, Cybernetics and Informatics SCI'97*, Caracas, Vénézuela, 1997, pp. 316–323.

[9] H. Eraqi and S. Abdelazeem, "An on-line Arabic handwriting recognition system based on a new on-line graphemes segmentation technique," in *Proceedings of the 11th International Conference on Document Analysis and Recognition*, Beijing, China, 2011, pp. 409-413.

[10] R. I. Elanwar, M. A. Rashwan, and S. A. Mashali, "Simultaneous segmentation and recognition of Arabic characters in an unconstrained on-line cursive handwritten document," in *Proceedings of World Academy of Science, Engineering and Technology (WASET)*, Germany, 2007, pp. 288-291.

[11] A. Elbaati, H. Boubaker, M. Kherallah, H. Elabed, A. Ennaji, and A.M. Alimi, "Arabic handwriting recognition using restored stroke chronology," in *Proceedings of the International Conference on Document Analysis and Recognition*, Barcelona, Espagna, 2009, pp. 411–415.

[12] L. Likforman-Sulem, A. Hanimyan, and C. Faure, "A hough based algorithm for extraction text lines in handwritten documents," in *Proceedings of the third International Conference on Document Analysis and Recognition*, Montreal, Canada, 1995, pp. 774–777.

[13] H. Boubaker, A. Elbaati, M. Kherallah, H. Elabed, and A.M. Alimi, "Online Arabic handwriting modeling system based on the graphemes segmentation," in *Proceedings of the 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 2061–2064.

[14] A. Hennig, N. Sherkat, and R.J. Whitrow, "Zone-Estimation for Multiple Lines of Handwriting Using Approximate Spline Functions," in *Proceedings of the 5th International Workshop Frontiers in Handwriting Recognition*, Colchester, UK, 1996, pp. 325-328.

[15] Z. Razak, K. Zulkiflee, M. Yamani, and I. Idris, "Off-line handwriting text line segmentation: a review," *International Journal of Computer Science and Network Security*, vol. 8, no. 7, pp. 12-20, Jul. 2008.

[16] M. A. Abuzaraida and A. M. Zeki, "Segmentation techniques for online Arabic handwriting recognition: a survey," in *Proceeding of 3rd International Conference on ICT4M*, Jakarta, Indonesia, 2010, pp. D37–D40.

[17] S. Izadi, M. Haji, and C. Y. Suen, "A new segmentation algorithm for online handwritten word recognition in persian script," in *Proceedings of the International Conference on Frontiers in*

Handwriting Recognition, Montréal, Canada, 2008, pp. 1140–1142.

[18] K. Daifallah, N. Zarka, and H. Jamous, "Recognition-based segmentation algorithm for on-line Arabic handwriting," in *Proceedings of the International Conference on Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 877–880.

[19] J. Sternby, J. Morwing, J. Andersson, and C. Friberg, "On-line Arabic handwriting recognition with templates," *Pattern Recognition Journal*, vol. 42, no. 12, pp. 3278-3286, 2009.

[20] M. Liwicki, A. Graves, H. Bunke, and J. Schmidhuber, "A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks," in *Proceedings of the 9th International Conference on Document Analysis and Recognition,* Curitiba, Brazil, 2007, pp. 367–371.

[21] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[22] M. Liwicki and H. Bunke, "Combining diverse on-line and off-line systems for handwritten text line recognition," *Pattern Recognition*, vol. 42, no. 12, pp. 3254–3263, 2009.

[23] J. Schenk, J. Lenz, and G. Rigoll, "Novel script line identification method for script normalization and feature extraction in on-line handwritten whiteboard note recognition," *Pattern Recognition Journal*, vol. 42, no. 12, pp. 3383–3393, 2009.

[24] A. M. Alimi, "Evolutionary neuro-fuzzy approach to recognize on-line Arabic handwriting," in *Proceedings of the 4th International Conference on Document Analysis and Recognition*, Ulm, Germany, 1997, pp. 382-386.

[25] N. Mezghani, A. Mitiche, and M. Cheriet, "Bayes classification of online Arabic Characters by Gibbs modeling of class conditional densities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1121-1131, Jul. 2008.

[26] M. Kherallah, L. Haddad, and A. M. Alimi, "On-line handwritten digit recognition based on trajectory and velocity modeling," *Pattern Recognition Letters*, vol. 29, no. 5, pp. 580–594, 2008.

[27] M. Kherallah, F. Bouri, and A. M. Alimi, "On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 1, pp. 153–170, 2009.

[28] M.Ma, D. W. Park, S. K. Kim, and S. An, "Online recognition of handwritten Korean and English characters," *Journal of Information Processing Systems*, vol. 8, no. 4, pp. 653-668, 2012.

[29] T. Al-Sheikh and S. El-Taweel, "Real-time Arabic handwritten character recognition," *Pattern Recognition*, vol. 23, no. 12, pp. 1323–1332, 1990.

[30] N. Mezghani, A. Mitiche, and M. Cheriet, "On-line recognition of handwritten Arabic characters using a Kohonen neural network," in *Proceeding of the International Workshop on Frontiers in Handwriting Recognition*, Niagara-On-the-Lake, Canada, 2002, pp. 490–495.

[31] N. Mezghani, A. Mitiche, and M. Cheriet, "Combination of pruned Kohonen maps for on-line Arabic characters recognition," in *Proceedings of the 7th International Conference on Document Analysis and Recognition*, Edinburgh, Scotland, 2003, pp. 900–905.

[32] A. M. Alimi, "Evolutionary computation for the recognition of on-line cursive handwriting," *IETE Journal of Research*, vol. 48, no. 5, pp. 385-396, 2002.

[33] H. Boubaker, M. Kherallah, and A. Alimi, "New strategy for the on-line handwriting modelling," in *Proceedings of the 9th International Conference on Document Analysis and Recognition*, Curitiba, Brazil, 2007, pp. 1233–1247.

[34] R. Plamondon, A. M. Alimi, "Speed/accuracy trade-offs in target-directed movements," *Behavioral and Brain Sciences*, vol. 20, no. 2, pp. 279–349, 1997.

[35] H. Boubaker, A. Chaabouni, N. Tagougui, M. Kherallah, and A. M. Alimi "Handwriting and hand drawing velocity modeling by superposing beta impulses and continuous training component," *International Journal of Computer Science Issues*, vol. 10, no. 5, pp. 57-63, 2013.

[36] Ch. C. Tappert, Ch. Y. Suen, and T. Wakahara, "The state of the art in on-line handwriting recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 12, no. 8. pp. 787–808, 1990.

[37] H. Boubaker, M. Kherallah, and A. M. Alimi, "New algorithm of straight or curved baseline detection for short Arabic handwriting writing," in *Proceedings of the 10th International Conference on Document Analysis and Recognition*, Barcelona, Espagna, 2009, pp. 778–782.

[38] G. Menier, G. Lorette, and P. Gentric, "A new modeling method for on-line handwriting recognition," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, Canada, 1995, pp. 499–503.

[39] M. Côté, M. Cheriet, E. Lecolinet, and C. Y. Suen, "Automatic reading of cursive scripts using human knowledge," in *Proceedings of the 9th International Conference on Document Analysis and Recognition*, Curitiba, Brazil, 2007, pp. 107–111.

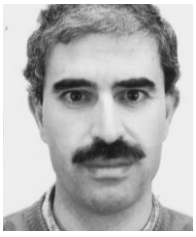
[40] P. Nagabhushan, S. A. Angadi, and B. S. Anami, "Geometric model and projection based algorithms for tilt correction and extraction of acsenders/descenders for cursive word recognition," in *Proceeding International Conference on Signal Processing, Communications and Networking*, Chennai, India, 2007, pp. 488–491.

[41] N. Tagougui, M. Kherallah, and A. M. Alimi, "Online Arabic handwriting recognition: a survey," *International Journal on Document Analysis and Recognition*, vol. 16, no. 3, pp. 209-226, 2012.

[42] A. Chaabouni, H. Boubaker, M. Kherallah, H. El-Abed, and A. M. Alimi, "Static and dynamic features for writer identification based on multi-fractals," *International Arabic Journal on Information Technologies*, vol. 11, no. 4, pp. 416-424, 2014.

[43] M. Hamdani, H. Elabed, M. Kherallah, and A. M. Alimi. "Combining multiple HMMs using on-line and off-line features for off-line arabic handwriting recognition," in *Proceedings of the 10th International Conference on Document Analysis and Recognition*, Barcelona, Espagna, pp. 201–205, 2009.

[44] Ph. Dreuw, S. Jonas, and H. Ney, "White-space models for offline Arabic handwriting recognition," in *Proceedings of the International Conference on Pattern Recognition*, Tampa, FL, 2008, pp. 1–4.



Houcine Boubaker

He was born in Kalaat El Andalous (Tunisia) in 1973. He graduated in Electrical Engineering in 1995, obtained a master degree in Systems Analyses and Digital Signal Processing in 1997. He is a researcher and a Ph.D. student in Electrical & Computer Engineering at the University of Sfax. His research interest includes trajectory modeling and applications of intelligent methods to pattern recognition. He focuses his research on drawing, Arabic handwriting and arm – hand movements modeling and Analyses. He is an IEEE student member and affiliate to the Research Group on Intelligent Machines laboratory (REGIM).

Najiba Tagougui

She was born in Sfax (Tunisia) in 1982. She graduated in Computer Sciences in 2005, obtained a master degree in News technologies of dedicated computer systems in 2007. She is now a Ph.D. student in Computer Systems Engineering at the University of Sfax. His research interest includes applications of intelligent methods to pattern recognition. She focuses her research on intelligent pattern recognition especially Arabic Handwriting Recognition. She is an IEEE student member and affiliate to the Research Group on Intelligent Machines laboratory (REGIM).

Haikal El Abed

He is a Ph.D. Senior Research Engineer at the Braunschweig Technical University, Germany. Since 2001, he has been working at the Institute for Communications Technology (IfN), Department of Signal Processing for Mobile Information Systems. He has specialized in image and signal processing, document analysis systems design and configuration, and Arabic/Latin manuscripts recognition. He coordinated different national and international research projects and is one of the developers of the IfN/ENIT–Database. He organized the Arabic Handwriting Recognition Competition at the ICDAR 2005, 2007 and 2009, He has more than 70 papers, including journal papers and book chapters. He is a member of IEEE, DAGM, IAPR (TC-10 and TC-11), and VDE/VDI and a frequent reviewer for international journals.

Monji Kherallah

He was born in Sfax (Tunisia) in 1963. He graduated in Electrical Engineering 1989, obtained a Ph.D. in Electrical Engineering in 2008. He is now a professor in Electrical & Computer Engineering at the University of Sfax. His research interest includes applications of intelligent methods to pattern recognition and industrial processes. He focuses his research on intelligent pattern recognition especially Arabic Handwriting Recognition. He is member of the editorial board of "Pattern Recognition Letters". He was a member of the organization committee of the International Conference on Machine Intelligence ACIDCA-ICMI'2005. He is an IEEE member and a frequent reviewer for international journals.

Adel M. Alimi

He was born in Sfax (Tunisia) in 1966. He graduated in Electrical Engineering 1990, obtained a Ph.D. and then an HDR both in Electrical & Computer Engineering in 1995 and 2000 respectively. He is now professor in Electrical & Computer Engineering at the University of Sfax. His research interest includes applications of intelligent methods (neural networks, fuzzy logic, evolutionary algorithms) to pattern recognition, robotic systems, vision systems, and industrial processes. He focuses his research on intelligent pattern recognition, learning, analysis and intelligent control of large scale complex systems. He is associate editor and member of the editorial board of many international scientific journals. He was guest editor of several special issues of international journals (e.g. Fuzzy Sets & Systems, Soft Computing, Journal of Decision Systems, Integrated Computer Aided Engineering, Systems Analysis Modeling and Simulations). He is an IEEE senior member.