Abstract—We present in this paper a new approach of online Arabic handwriting modeling based on the graphemes segmentation. This segmentation rests on the previous detection of baseline. It involves the detection of two types of topologically meaningful points: the backs of the valleys adjoining the baseline and the angular points. The stage of features extraction allows to model the shapes of segmented graphemes by relevant geometric parameters and to estimate their diacritics fuzzy affection rates. The test results show a significant improvement in recognition rate with the introduction of new pertinent parameters.

Keywords—online handwriting; baseline detection; grapheme segmentation; handwriting modeling

I. INTRODUCTION

The cursive or semi-cursive handwriting such as Arabic or Latin, represent concatenations of a limited number of basic graphic shapes called graphemes. The graphemes can represent characters or pseudo-characters. Their cursive sequence verifies some dynamic properties and topologic rules related to the linearity and interconnection [15, 17, 18]. In another sense, these rules can serve to segment the cursive script in its basic components: the graphemes.

The graphemes segmentation is an essential step in an analytic recognition process of cursive handwriting in the context of an extended or infinite lexicon [1, 5, 6]. Indeed, the extraction of parametric or structural characteristics of segmented graphemes can detect the basic forms of the script to recognize [12, 13]. The previous detection of the baseline allows to detect the topologically special points which limit the graphemes: the bottom of the valleys close to the baseline and the angular points. The algorithm that we developed consists of three modules: the baseline detection, the graphemes segmentation and features extraction. We will present, successively in the three following sections of this paper the different modules of the algorithm before ending up presenting tests and results.

II. BASELINE DETECTION MODULE

The baseline detection is an essential stage in a grapheme segmentation process of a cursive or semi-cursive handwriting [4, 5, 7, 8, 9].

The developed baseline detection process consists of two stages: The first one, being a basic stage, permits the detection of the points regrouping of aligned neighbourhood [16]. For this, we inspect the alignment and the tangent direction accordance of each current point \(M_i\) according to the elements of the points regroupings to which it is a candidate element, using two criteria:

- **Validation criterion:**
  A point candidate \(M_k\) can be assigned to the points regrouping \([M]_n\) if it verifies (1):

  \[
  \forall M_{n,i} \in [M]_n \text{ we have } \Delta\alpha_{i,k} + \Delta\alpha_{k,i} < \Delta\alpha_{\lim} \tag{1}
  \]

  With \(\Delta\alpha_{\lim}\) is the tolerance limit of the absolute deviation angles between the trajectory tangents. And:

  \[
  \Delta\alpha_{i,k} = \alpha_{tg M_{n,i}} - \alpha_{tg M_{n,j}} \quad \text{the slant angle of the direction} \quad (M_{n,i}, M_k)
  \]

  \[
  \Delta\alpha_{k,i} = \alpha_{tg M_{n,j}} - \alpha_{tg M_{n,i}} \quad \text{the slant angle of the direction} \quad (M_{n,i}, M_k)
  \]

  (See Fig. 1)

Figure 1. Verification of the trajectory neighborhoods alignment.
- **Affectation criterion:**

A point candidate $M_k$ verifying the validation conditions (1) to several regroupings $\{M\}_{k=1,q}$, is assigned to the regrouping of index $m$: $\{M\}_m$ where agrees best its trajectory tangent direction with those of the other members as well as with the directions of interpolation $(M_k, M_{m,i})$ in accordance with the following criterion (2):

$$\Delta \theta_{M_k}(m) = \min_{n=1,...,q} \left\{ \Delta \theta_{M_k}(n) \right\}$$

with:

$$\Delta \theta_{M_k}(n) = \frac{1}{N_n} \sum_{M_{n,i} \in \{M\}_n} \left( \Delta \alpha_{n,i} + \Delta \alpha_{k,k} \right)$$

(2)

Where $N_n$ is the initial size of the $\{M\}_n$ regrouping and $m \in \{1,...,q\}$.

A new points regrouping is initialized when the point candidate $M_k$ is not included in any already constituted regrouping.

The baseline detection, at this stage of the treatment, consists in looking for the most numerous regrouping among the points regroupings that are constituted (see Fig. 2).

- The average angle $\theta_{m,\text{curv}}$ of graphemes absolute curvature.
- The bending on the left ($bbl$) of the barycentre of the set of contact points between segmented graphemes and baseline (see Fig. 3).

The function of assessment $S$ that takes into account these different parameters is expressed by the following formula (3):

$$S = (\alpha_1 \cdot \text{npt}) \cdot (\alpha_2 \cdot \theta_{\text{bbl}}) \cdot (\alpha_3 \cdot \theta_{\text{curv}}) \cdot (\alpha_4 \cdot \text{bbl})$$

(3)

In order to estimate correctly the weighting coefficients $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ we assimilated the $S$ function to the output of an ADALINE network simple layer trained according to the 'least mean square error' rule.

Fig. 4 shows the result of the correction step of the baseline detection error obtained in Fig. 3.

### III. GRAPHEMES SEGMENTATION

A grapheme is a distinctive unit of the handwriting that represent a whole character or a section of its tracing. Example: several Arabic characters as ‘‘’, ‘‘’, ‘‘’, ‘‘’ include one or several graphemes named 'nabra' ‘‘.’

The segmentation of the pseudo - words in graphemes is based on the detection of two typographically significant points [19] (see Fig. 5):

- The bottom of the valleys: the point of an inter - grapheme ligature adjoining the baseline with a horizontal tangent.
- The angular points: the extremum point of a trajectory turn back.

### IV. GRAPHEMES MODELING

The objective of this module consists in extracting relevant parametric features that characterize each element of basics graphemes which constitute Arabic handwriting [2, 3]. We associate a bounding box and reference points for each
segmented grapheme as explained in following paragraphs (see Fig. 6).

A. The measurements of the bounding box

The letters or the Arabic graphemes can be partially characterized by their measurements (height and width). For example, the graphemes '١' and 'ﺐ' are quite distinct considering only the dimensions of their bounding box [10].

B. The relative position of the bounding box

The vertical relative position of the bounding box permits to discriminate three sets of graphemes. Indeed according to their positions respect to the baseline, we distinguish the graphemes that are written in over of the baseline, of others that descend underneath the baseline and the diacritics.

C. The positions of the reference points

The three considered points reference marks are :

- The starting point of the grapheme trajectory \( M_1 \).
- The point of arrival \( M_n \).
- The point corresponding to the absolute minimum of curvature radius \( M_i \in \left] M_1, M_n \right] \) (see Fig. 8 b/).

The positions of the points reference marks, \( M_1, M_n, M_i \) in the bounding box give a preview on the shape of the grapheme trajectory. These positions are defined in respect to the left lower summit of the bounding box in the horizontal and vertical direction by the ratios \( R_H \) and \( R_V \).

D. Direction of the trajectory on the level of the reference points

In the objective to get more precision for the trajectory model, we determine the slant angles \( \theta_1, \theta_n \) and \( \theta_m \) of the tangent to the trajectory respectively to the three reference points \( M_1, M_n \) and \( M_n \) (see Fig. 8 b/).

E. Grapheme curvature features

In order to study the trajectory curvature direction of the grapheme, we measure its continuous \( \alpha_{Ca} \) and absolute curvature angles \( \alpha_{Aa} \) along the tracing:

\[
\alpha_{Ca} = \sum_{i=2}^{n} \left( \theta_{M_i} - \theta_{M_{i-1}} \right) = \text{continus } \theta_n - \text{continus } \theta_1
\]

\[
\alpha_{Aa} = \sum_{i=2}^{n} \left| \theta_{M_i} - \theta_{M_{i+1}} \right|
\]

F. Diacritics detection and fuzzy affectation

Statistics made on handwriting strokes extracted from normalized samples of the ADAB database permitted to define the dimension and position thresholds that distinguish diacritics from main graphemes (see Fig. 8 a/). Then we built a fuzzy estimator for a proportional affectation of diacritics to each main grapheme.

In the input, the estimator gets the membership degree \( \mu_{\text{diac.1} / G_j} \) of the \( i \)th diacritic centroid to the \( j \)th grapheme \( G_j \), and the parameters \( \Delta X_i, \Delta Y_i \) which define respectively the horizontal and vertical dimensions of the \( i \)th diacritic (see Fig. 7). In the output we get the proportional rate \( T_{a_{\text{diac.top}}} / G_j \) of fuzzy affectation of the \( i \)th diacritic to the \( j \)th main grapheme. Then for each main grapheme \( G_j \), we estimate tow total fuzzy rates ; \( T_{a_{\text{diac.top}}} / G_j \) and \( T_{a_{\text{diac.down}}} / G_j \) respectively for the top and the down diacritics affectation (see Fig. 8 b/) by the following formulas :

\[
T_{a_{\text{diac.top}}} / G_j = \sum_{i=1}^{\text{numbe of Top diacritics}} T_{a_{\text{diac.top}}} / G_j
\]

\[
T_{a_{\text{diac.down}}} / G_j = \sum_{k=1}^{\text{numbe of Down diacritics}} T_{a_{\text{diac.down}}} / G_j
\]

Figure 7. Estimation of the diacritics membership to the main graphemes.

In the input, the estimator gets the membership degree \( \mu_{\text{diac.1} / G_j} \) of the \( i \)th diacritic centroid to the \( j \)th grapheme \( G_j \), and the parameters \( \Delta X_i, \Delta Y_i \) which define respectively the horizontal and vertical dimensions of the \( i \)th diacritic (see Fig. 7). In the output we get the proportional rate \( T_{a_{\text{diac.top}}} / G_j \) of fuzzy affectation of the \( i \)th diacritic to the \( j \)th main grapheme. Then for each main grapheme \( G_j \), we estimate tow total fuzzy rates ; \( T_{a_{\text{diac.top}}} / G_j \) and \( T_{a_{\text{diac.down}}} / G_j \) respectively for the top and the down diacritics affectation (see Fig. 8 b/) by the following formulas :
The statistics show that for Arabic handwriting, diacritics are often shifted to the left of the main grapheme which explains the asymmetric shape chosen for the membership function $\mu_{\text{diac}, j/G_j}$ (see Fig. 7).

V. TESTS AND RESULTS

In the evaluation phase, we applied the system on the online database ADAB of Tunisian names towns using the HMM Tool Kit ‘HTK’ as classification module [20]. We obtained the following recognition results for three ameliorated versions of the system (Tab.1):

<table>
<thead>
<tr>
<th>Version</th>
<th>Recognition Rate (Top 1)</th>
<th>Recognition Rate (Top 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 1</td>
<td>57.87</td>
<td>72.89</td>
</tr>
<tr>
<td>Version 2</td>
<td>79.46</td>
<td>93.58</td>
</tr>
<tr>
<td>Version 3</td>
<td>87.13</td>
<td>98.04</td>
</tr>
</tbody>
</table>

We note the successive improvement of the of recognition rates in top1 and top2 with the adjustment of the filters and the fuzzy diacritics modeling.

VI. CONCLUSION

We presented in this paper an online Arabic handwriting modeling system based on graphemes segmentation. The system consists of three modules: detection of the baseline, graphemes segmentation and features extraction. The method developed in the first module is characterized by the consideration of geometrical and topological features for the baseline detection and correction. In the second module, we use the detected baseline to look for particular points: the bottom of the valleys and the angular points for the segmentation of the cursive handwriting trajectory in graphemes. Finally the third module extracts parameters to models the position, the shape, and the fuzzy affectation rate of diacritics associated to each segmented grapheme. The test results show a significant improvement in recognition rate with the introduction of new pertinent parameters.

ACKNOWLEDGMENT

The authors acknowledge the financial support of this work by grants from the General Direction of Scientific Research and Technological Renovation (DGRST), Tunisia, under the ARUB program 01/UR/11/02.

REFERENCES


