

Online Arabic Handwriting Modeling System based on the Graphemes Segmentation

Houcine Boubaker , Abdelkarim El Baati ,
Monji Kherallah , Adel M. Alimi

REGIM: REsearch Group on Intelligent Machines
University of Sfax

National School of Engineers (ENIS)
BP 1173, Sfax, 3038, Tunisia

{Houcine-boubaker , abdelkarim.elbaati ,
monji.kherallah , adel.alim}@ieee.org

Haikel Elabed

Institute for Communications Technology (IfN)
Technische Universität

D-38106 Braunschweig, Germany
elabed@tu-bs.de

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 affectation 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_k 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):

$$\forall M_{n,i} \in \{M\}_n \text{ we have } \Delta\alpha_{i,k} + \Delta\alpha_{k,i} < \Delta\alpha_{lim} \quad (1)$$

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

$$\Delta\alpha_{i,k} = \left| \alpha_{tg M_{n,i}} - \text{the slant angle of the direction } (M_{n,i}, M_k) \right|$$

$$\Delta\alpha_{k,i} = \left| \alpha_{tg M_{n,j}} - \text{the slant angle of the direction } (M_{n,i}, M_k) \right|$$

(See Fig. 1)

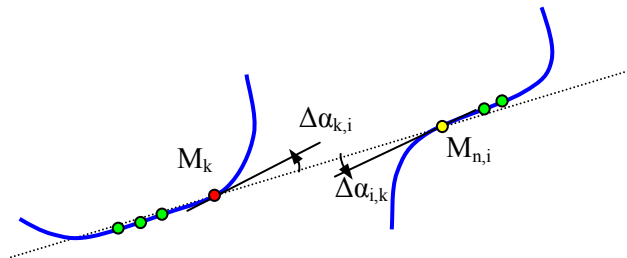


Figure 1. Verification of the trajectory neighborhoods alignment.

- Affection criterion:

A point candidate M_k verifying the validation conditions (1) to several regroupings $\{M\}_{1,\dots,q}$, 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) :

$$\Delta\theta_{M_k}(m) = \text{Min}_{n=1,\dots,q} \{ \Delta\theta_{M_k}(n) \}$$

$$\text{with : } \Delta\theta_{M_k}(n) = \frac{1}{N_n} \cdot \sum_{M_{n,i} \in \{M\}_n} \{ \Delta\alpha_{k,i} + \Delta\alpha_{i,k} \} \quad (2)$$

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

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).

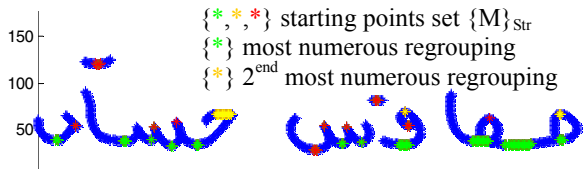


Figure 2. constitution of the points regroupings.

The examination of the baseline detection errors shows that they are classified in two cases [16]:

- Confusion of the baseline with the lower limit line for the cases of words composed essentially or exclusively of isolated character or of legs as 'ر', 'و', 'ز', 'ن' (example see Fig. 3) .
- Confusion of the baseline with the median zone line, or the superior limit line, due to the writing style or to the presence of particular calligraphic effects.

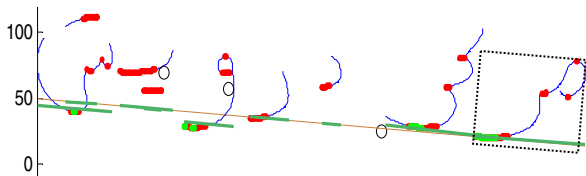


Figure 3. Examples of baseline detection errors.

To discern and to treat the baseline detection errors we opted for a function of assessment considering the first three most extended regrouping in order to optimize the detection result. This cost function excels the size of the points regrouping (npt) and penalize:

- The average angle θ_{\cap_bl} of intersection between the upward trajectory and baseline.

- The average angle $\theta_{m_|Curv|}$ of graphemes absolute curvature.
- The bending on the left (bb_l) 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) :

$$S = (\alpha_1 \cdot npt) - (\alpha_2 \cdot \theta_{\cap_bl}) - (\alpha_3 \cdot \theta_{m_|Curv|}) - (\alpha_4 \cdot bb_l) \quad (3)$$

In order to estimate correctly the weighting coefficients $\alpha_1, \alpha_2, \alpha_3,$ and α_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.

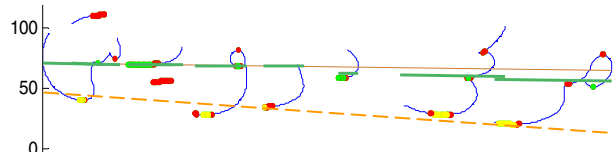


Figure 4. Example of baseline correction (green).

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 'سا', 'ب', 'ت' include one or several graphemes named 'nabra' 'ا'.

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.

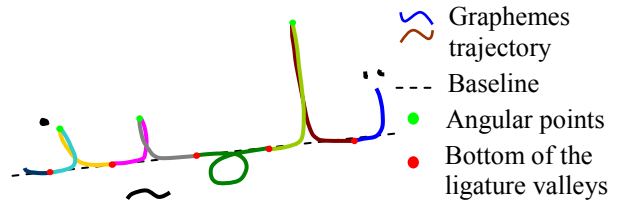


Figure 5. The typographically significant points and graphemes segmentation.

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 'ا' and 'ب' are quite distinct considering only the dimensions of their bounding box [10].

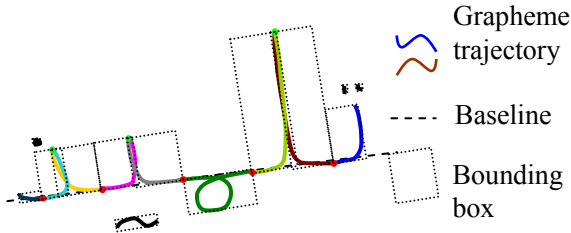


Figure 6. The bounding box of segmented graphemes.

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 $M_i \in]M_1, M_n[$ (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 $\theta_1, \theta_i,$ and $\theta_n,$ of the tangent to the trajectory respectively to the three reference points $M_1, M_i,$ 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 α_{Ca} and absolute curvature angles α_{Aa} along the tracing:

$$\alpha_{Ca} = \sum_{i=2}^n (\theta_{M_i} - \theta_{M_{i-1}}) = \text{continus_}\theta_n - \text{continus_}\theta_1 \quad (4)$$

$$\alpha_{Aa} = \sum_{i=2}^n |\theta_{M_i} - \theta_{M_{i-1}}|$$

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.

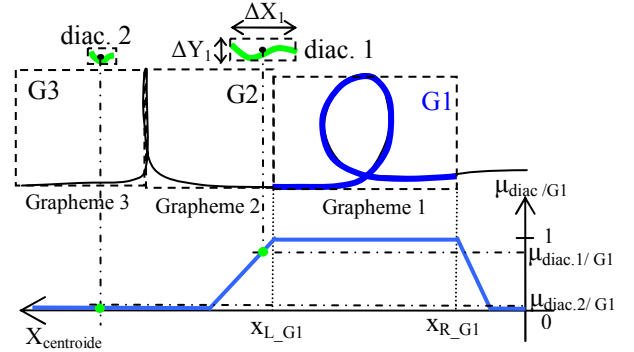


Figure 7. Estimation of the diacritics membership to the main graphemes.

In the input, the estimator gets the membership degree $\mu_{\text{diac}_i / G_j}$ 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 $T_{a_{\text{diac}_i / G_j}}$ 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 :

$$T_{a_{\text{diac_top} / G_j}} = \frac{\text{number of Top diacritics}}{\sum_i T_{a_{\text{diac}_i / G_j}}} \quad (5)$$

$$T_{a_{\text{diac_down} / G_j}} = \frac{\text{number of Down diacritics}}{\sum_k T_{a_{\text{diac}_k / G_j}}}$$

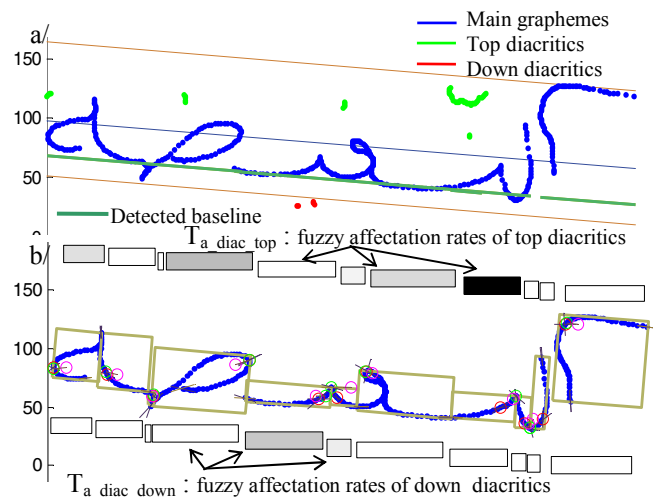


Figure 8. a/ Diacritics detection using tracing dimension and position respect to the baseline b/ Estimation of the fuzzy rates of top and down diacritics affectation to each main grapheme.

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 $\mu_{\text{diac}_i/\text{G}_j}$ (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):

Version 1: without diacritics treatment (Icdar 2009 competition) [11].

Version 2: after adjusting the filters and without diacritics treatment.

Version 3: after adjusting the filters and with the extraction and fuzzy affectation of diacritics.

TABLE I. RECOGNITION RATE OBTAINED ON THE ADAB DATABASE SET 1 AND 2

System version	ADAB Set 1		ADAB Set 2	
	Top 1	Top 5	Top 1	Top 5
Version 1	57.87	72.89	54.26	66.38
Version 2	79.46	93.58	77.61	89.72
Version 3	87.13	98.04	84.79	97.45

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

[1] A. M. Alimi, “Evolutionary computation for the recognition of on-line cursive handwriting,” IETE Journal of Research, Volume 48, Issue 5 SPEC., September 2002, pp 385-396.

[2] A. M. Alimi, “Neuro-fuzzy approach to recognize Arabic handwritten characters,” Proceedings of the IEEE International Conference on Neural Networks, ICNN Volume 3, 1997, Pages 1397-1400

[3] A. M. Alimi, “Evolutionary neuro-fuzzy approach to recognize on-line Arabic handwriting,” Proceedings of the International Conference on Document Analysis and Recognition, ICDAR Volume 1, 1997, Pages 382-386

[4] M. Pechwitz, V. Märgner. “Baseline Estimation For Arabic Handwritten Words”. Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR’02), 2002, pp 479 – 484.

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

[6] A. Amin. “Off-line Arabic character recognition: The state of the art.” Pattern Recognition, 1998, pp 517–530.

[7] M. Côté, M. Chériet, C. Suen and E. Lecolinet (1996), “Détection des Lignes de Base de Mots Cursifs à l'aide de l'Entropie,” Colloque sur l'Intelligence Artificielle dans les Technologies de l'Information, Montréal Canada, may 1996.

[8] Z. Razak, K. Zulkiflee, M. Yamani, I. Idris. “Off-line Handwriting Text Line Segmentation: A Review.” IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008.

[9] Likforman-Sulem L., A. Hanimyan, C. Faure (1995) “A Hough based algorithm for extracting text lines in handwritten documents,” In Proc. of the Third Int. Conference on document analysis and recognition (ICDAR), Montreal, Canada, August 1995, pp. 774-777.

[10] H. Miled, C. Olivier, M. Chériet, K. Romeo-Pakker, “Une Méthode Rapide de Reconnaissance de l'Écriture Arabe Manuscrite,” sixth symposium GRETSI - September 1997 – Grenoble.

[11] E. Grosicki, H. El Abed, ICDAR 2009 “Handwriting Recognition Competition,” ICDAR10th 2009 Barcelona pp 1398 – 1402.

[12] M. Côté, M. Chériet 2, E. Lecolinet and C.Y. Suen, “Automatic Reading of Cursive Scripts Using Human Knowledge,” ICDAR 97 pp 107-111.

[13] M. Kherallah, L. Haddad, and A. M. Alimi, “On-line handwritten digit recognition based on trajectory and velocity modeling,” Pattern Recognition Letters Volume 29, Issue 5, 1 April 2008, Pages 580-594

[14] H. Zouari, L. Heutte, Y. Lecourtier, A. M. Alimi, “Building diverse classifier outputs to evaluate the behavior of combination methods: The case of two classifiers,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) Volume 3077, 2004, Pages 273-282

[15] R. Plamondon, A. M. Alimi, “Speed/accuracy trade-offs in target-directed movements,” Behavioral and Brain Sciences Volume 20, Issue 2, 1997, Pages 279-349

[16] H.Boubaker, M. Kherallah, and A. M. Alimi , “New Algorithm of Straight or Curved Baseline Detection for Short Arabic Handwritten writing,” The 10th International Conference on Document Analysis and Recognition ICDAR 2009, Barcelona - Espagna

[17] H. Boubaker, M. Kherallah, and A. Alimi, “New Strategy for the On-Line Handwriting Modelling,” Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) vol. 2, pp.1233-1247

[18] H. Boubaker, M. Kherallah, and A. M. Alimi , “Une nouvelle stratégie pour l'extraction des caractéristiques de l'écriture manuscrite en ligne,” Conférence Internationale Francophone sur le traitement Electronique des Documents CIFED 2006, Fribourg,- Fédération Suisse, Switzerland

[19] A. Elbaati, H. Boubaker, M. Kherallah, H. Elabed, A. Ennaji, and A.M. Alimi. “Arabic handwriting recognition using restored stroke chronology,” In International Conference on Document Analysis and Recognition ICDAR 2009.

[20] M. Hamdani, H. Elabed, M. Kherallah, and A.M. Alimi. “Combining multiple hmms using on-line and o_-line features for o_-line arabic handwriting recognition,” In International Conference on Document Analysis and Recognition ICDAR 2009.