3rd International Conference on Automation, Control, Engineering and Computer Science (ACECS'16)

Proceedings of Engineering & Technology (PET)
pp. 46-51
Copyright IPCO-2016

Online Arabic Character Recognition Using Global And Local Features

Houda Nakkach, Sofiene Haboubi, and Hamid Amiri
LR-11-ES17 Signal, Images et Technologies de l'Information $LR-SITI-ENIT$
Université de Tunis El Manar, Ecole Nationale d'Ingénieur de Tunis 1002, Tunis Le Belvédère, Tunisie
Email: h_nakkach@yahoo.fr, sofiene_haboubi@yahoo.fr, and hamid.amiri@enit.rnu.tn

Abstract—**In this paper, we present a new method to describe online Arabic characters. This is based on a vector of sequence sets features per point to be fed to classifier. After the pre-processing step, the x and y coordinates of online signal are used to extract a some of local and global features such as : direction, curvature, down or up of the pen, normalized chain code and the Fourier descriptors. These features are then fed in the Support Vector Machine (SVM) for classification. The method of representation based on global and local features achieved accuracy of about 92.43% with SVM classifier.**

I. Introduction

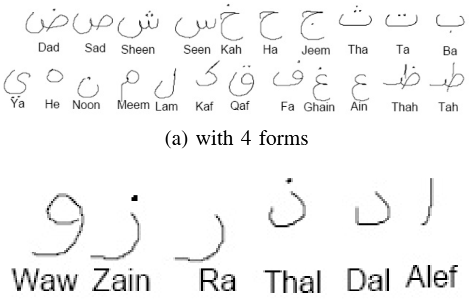
The researches in the field of signal processing line have extended considerably in recent years. Because of the complex problems in the recognition systems, which arise mainly by stylus-oriented interfaces that seek to integrate handwriting as a new human-computer interaction modality. In this acquisition mode, the handwriting recognition is often done at the same time as the writer is writing. The acquired information is to follow the path of the pen tip on the tablet, which is stored as a dependent signals of the time, i.e. a sequence of coordinates of points ordered in time (x (t) , y (t)). Whether recognition of online or offline handwritten forms, the related issues are particularly complex. In fact, there are two important steps related to this field: Features extraction and recognizer phases. The latter was the subject of most studies on the assumption that the classification was essential and should focus on this complex task. However, despite the complexity of the problem of recognizer class, a good representation results in a good classifier.

In this context, the aim of this work is to study the feasibility to mixture most approaches of features extraction presented in the literature and try to develop an approach to characterization of Arabic characters based on Fourier descriptor and chain codes. This proposal is based on a thorough review of existing approaches to introduce a method of features extraction to online Arabic characters; this method combines static and structural approaches.

II. SPECIFICITIES OF ARABIC WRITING

A. Character position in a pseudo-word

The Arabic alphabet consists in 28 letters whose forms vary depending on the position in the word. The letters are spelled differently depending on whether they are isolated, at the beginning, middle or end of the word. We distinguish 22 letters from the alphabet (Fig.1.a) with 4-forms of writing. The remaining six (Fig.1.b) may be attached to their successors, and so they have only two forms.



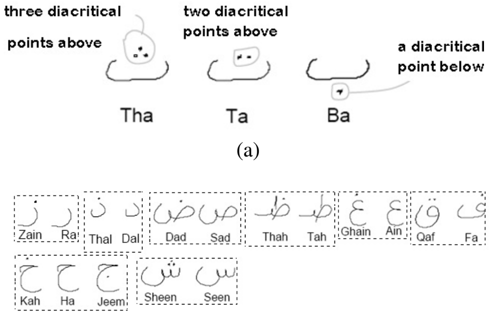
(a) with 4 forms

(b) with 2 forms

Fig. 1. alphabet forms of writing

B. Point diacritical

More than half of Arabic characters include in their forms diacritical points (1, 2 or 3). These points may be located above or below the character, but never in the top and bottom simultaneously. Multiple characters can have the same body but a number and / or position of various diacritics (fig.2 (a,b)). So, the Arabic alphabet consists in 28 letters from 18 different shapes and diacritical points. If the letter "Fa" is equated with the letter "Qaf", which is distinguished, only by their position relative to the write line (fig.3). Information that is not generally available for isolated characters was 17 distinct forms. Throughout our work, we refer to the 17 forms.


three diacritical points above
two diacritical points above
a diacritical point below
Tha
Ta
Ba
Zain
Ra
Thal
Dal
Dad
Sad
Thah
Tah
Ghain
Ain
Qaf
Fa
Kah
Ha
Jeem
Sheen
Seen

(a)

(b)

Fig. 2. Characters with same body

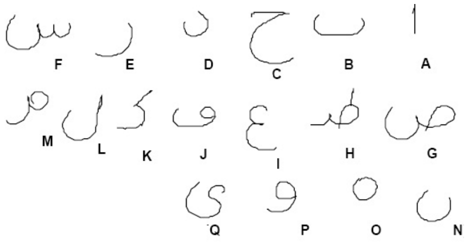


Fig. 3. 17 Forms of isolated Arabic characters

C. The variation in the size

The size of the Arabic characters may vary from one writer to another and even within the writing of a single writer. This can cause instability of the parameters at the stage of primitives. Many works have been proposed in the literature for normalizing the size of the Latin and Arabic writing, however none of them are accepted or regarded as a standard method. In [1], Madhvanath presented a method for the Latin script, this technique is based on the estimation of the average width characters.

III. FEATURES EXTRACTION

The objective of the features extraction step is to definition and selection of a set of features, for the system of recognition of cursive handwriting. This is a delicate and important work for the next recognition step. The choice of representation of the signal must write integrate various criteria, such as the complementarities of the selected features. The representation of the most common signal line is the time sequence of feature vectors, local measurements at each point of the normalized height of the path baseline.

In this regard, there are different approaches based on local or global features.

In [2], [3] and [4] Guyon, Schenkel and Poisson and all presented an approach that resample any signal into fixed number of points by interpolating methods plot and position relative to its center of gravity. The main features extracted at each point of the plot are the coordinates, the orientation of the path relative to the horizontal curvature and information about down or up to the pen.

In [5], Parizeau uses a transformation 2D space of the characters to a new region based fuzzy vector space (specifically a division into 6 zones) where seven fuzzy features are extracted. The first three measures characterize the curvature depending on its value: straight line (R), positively curved line (C +) and negatively curved line (C-). The other four relate to the orientation of the plot horizontal (H) Vertical (V), positive or negative oblique (O + or O-). The signal is then represented by a fuzzy vectors concatenation region coupled to each of the horizontal and vertical measurements of density for each region.

This approach is classified among approaches which consist of extracting a fixed number of characteristics of write signal. According to [3] and [6], Schenkel and Jaeger derived purely local characteristics, semi local or global from an initial sequence to address the problems of letters with different sizes. In [4] and [7], Alleau (for online domain) and Gnter were interested in checking the interest and the complementary characteristics, and they consider the number of features to select a fundamental step. On Besides, some studies seek to characterize the signal structural information such as the Freeman coding, directional coding of the path for 16 directions [8], encoding 36 elementary strokes [9] [10](see fig.4). As part of recognition system online, different authors have studied the contribution and complementarity of the static representation of the track relative to the dynamics of this signal. In [4] and [11], Poisson and Alimoglu present an approach combining both types of information to improve recognition performance. According to [12] the related attributes are divided into two groups: the stable mode (diacritical points and relative position of these points compared to the base line) and unstable (number of second segments and inclinations of the segments). Then, to obtain accurate performance of recognition, functions both quantitative (statistical / numerical) and qualitative (structural / topological) were defined and used in research at AOCR systems (Arabic Optical Character Recognition). The most commonly functions used are:

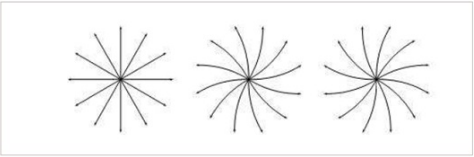


Fig. 4. 6 Sets of 36 basic strokes used to represent two-dimensional shapes [9], left to right: 12 tracks Rights, 12 convex, 12 concave

- Statistical approach
  This approach is frequently used to design systems for pattern recognition. In fact, its based on the statistical analysis of measurements made on forms to be recognized. It has also quantitative characteristics that represent the character.
- Structural approach
  This type of approach has qualitative characteristics that represent the structure of the character or stroke.

IV. RELATED WORKS

For related work in the literature, we consider the work of [4], [13], [14], which have a characterization method based on the mixture between statistical and structural approach(with tracking).

A. Mixture of Fourier descriptor and tangents

In [14], N. Mezghani presents the combination of two memories Kohonen driven on different representations of characters: the Fourier descriptors and the tangents to measure sample points sighted pretreatment. These two representations are complementary: the first is global, but the second is local. The performed tests have shown that the combination has significantly improved the recognition rate, compared to the base system consisting of a single memory.

B. Extraction of geometric characteristics of online signal

In [4], Poisson works on the recognition of handwritten Arabic script online as it defines a system that receives an input online signal (just the file of point coordinates of a character Arabic format). This signal is then resampled by a number of fixed points and normalized. To this standard signal, it extracts a sequence of vectors of the number of features per point to be presented to a classifier.

The work consists on the extraction sequence of points from local geometric information such as the direction of movement and the curvature of the trajectory in order to obtain a sequence of 7 features vector: coordinates, direction, curvature, down or up the pen - per point. subsectionMixture of Fourier descriptor and chain codes of freeman In [13], Rajput presents a new method for automatic recognition of handwritten character Marathi isolated offline. For the features extraction, he combined Fourier descriptor and standard chain codes. These characteristics are then introduced to a Support Vector Machine classifier (SVM).

V. OUR APPROACH

The aim of this article is to explore a new method based on combination of a statistic approach and a structural approach, by mixture of Fourier descriptors with the chain codes. Our method will be presented below in two steps.

A. Global Approach

1) Fourier Descriptor: We use Fourier descriptors as computed in Kuhl and Gardinas [15]. A closed contour (x(t),y(t)), t=1 …m, is approximated using N elliptic Fourier descriptors as :

$$\hat{X}(t) = A_0 + \sum_{i=1}^{n} a_n \cos\frac{2n\Pi}{L} + b_n \sin\frac{2n\Pi}{L} \quad (1)$$

$$\hat{Y}(t) = C_0 + \sum_{i=1}^{n} c_n \cos\frac{2n\Pi}{L} + d_n \sin\frac{2n\Pi}{L} \quad (2)$$

Where L is the contour length, $\hat{x}(t) \equiv x(t)$ and $\hat{y}(t) \equiv y(t)$ in the limit when $n \to \infty$. Coefficients an, bn, cn and dn, which are used as features are:

$$a_n = \frac{1}{L}\int_0^L x(t)\cos\frac{2n\Pi(t)}{L}dt \quad (3)$$

$$b_n = \frac{1}{L}\int_0^L x(t)\sin\frac{2n\Pi(t)}{L}dt \quad (4)$$

$$c_n = \frac{1}{L}\int_0^L y(t)\cos\frac{2n\Pi(t)}{L}dt \quad (5)$$

$$d_n = \frac{1}{L}\int_0^L y(t)\sin\frac{2n\Pi(t)}{L}dt \quad (6)$$

2) Challenges of the Global Approach: Fourier Descriptors present comprehensive and specific features for Arabic characters. this approach is considered insufficient for the recognition step. Thus, it must improve its combination with a method that goes into specific details of the Arabic characters, such as the detection directions.

B. Local approach

In the local approach, we adopt essentially the codes of the chain. This phase contains two main steps: the first is based on monitoring the trajectory of the pen to generate a vector V_0 to 7 features. The second step is based on the codes of the chain freeman, in order to generate a vector V_1 that will be transformed into more optimal vector V_2 which in turn will be transformed into vector V_3 that normalizes the write size. At the end, we concatenate the vector V_2 with V_3 in the vector V_4 with 16 features, which will be combined with V_0 to finally have V_5 with 23 features.

1) Approach with followed: The characters are entered on a digital tablet. They are represented as a sequence of coordinates[x(t),y(t)]. We chose to preserve the nature sequential of the information acquired by the tablet for a constant number of points regularly spaced along the path of the pen which provides a better recognition rate. At this level, the essential points or parameters to score are :

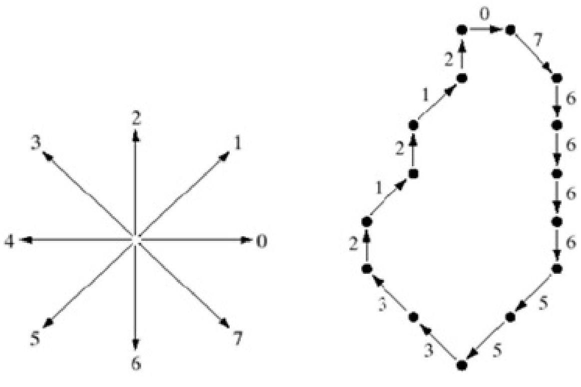
- Acquisition start: start writing and put pen.
- Coordinates: horizontal position x(t) , the vertical position y(t), the pressure of the pen on the tablet and the time t. Detect the pen status. The information of up and put pen (Pen_Up and Pen_down) characterizes a new feature or a new point at the character level . We specify the state of the pen in point n, as follows: PenUpDown = 0 and PenUpDown = 1.
- Then in order to facilitate the task of classifier, the extraction of points sequence of local geometric information such as the movement direction and the curvature of the path.

This produces a sequence of 7 features vectors: coordinates, direction, curvature, PenDown or PenUp by point.

We fix the number of points to Nmax for each character. In each of these points, a vector V_0 to 7 features is extracted. It contains the x and y coordinates, the directors cosines of the direction ($\cos\theta$, $\sin\theta$) and curvature ($\cos\Phi$, $\sin\Phi$) trajectory and the state of the pen at this point.

$$V_0\,[X(t), Y(t), cos(\theta), sin(\theta), cos(\Phi), sin(\Phi), PenUpDown] \tag{7}$$

2) Chain Code: The chain code of Freeman [16] are generated using the location of a boundary pixel, also called the starting pixel and then moving along the delimited zone or the counterclockwise direction, to find the next boundary pixel and assign the new pixel according to a code of the previous pixel location. The search process is achieved when the next pixel starting pixel is encountered. Codes can be 4 or 8-steering direction according to 4 - or 8-connectivity of a pixel to its neighbor contour pixel. A chain (8-management) coded image is given in fig.5. Chain code: 076666553321212



(a) (b)

Fig. 5. 8- connectivity chain codes

The approach adopted in this step:

1) Generate an initial vector V_1
We assume the chain code generated to present the outline of the shape shown in Fig.6, by browsing in counter clockwise as direction. (We fix Nmax to 50 points)

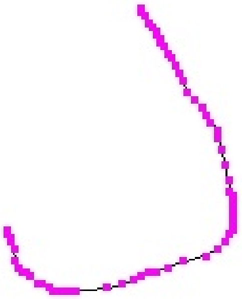


Fig. 6. ”Da” character

$$V_1 = [67777676777777777777767777676666650505605050550505 \tag{8}$$

2) Calculate the frequency of codes 0, 1, 2 . . . 7.
For the vector V_1: we have the following vector of V_2 frequency:

$$V_2 = [8000081121] \tag{9}$$

3) The frequency normalization is represented by the vector V_3, which is calculated by the following formula:

$$V_3 = \frac{V_2}{|V_2|} \tag{10}$$

For the previous example, we have:

$$V_3 = [0.160.00.00.00.00.160.220.43] \tag{11}$$

4) Concatenate the vectors V_2 and V_3 in V_4 size 16.
For the previous example, we have:

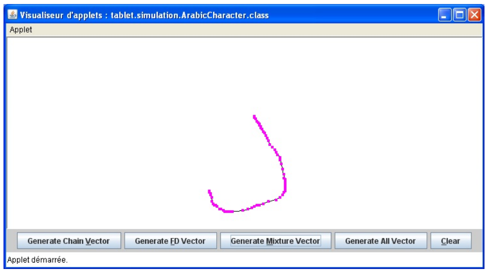
$$V_4 = [8.00.00.00.00.08.011.021.00.160.00.00.00.00.160.220.43 \tag{12}$$

We note that this representation is quite comprehensive and does not enter into the details of the components online character. It does not present the dynamic aspect and the variation in time of the signals that characterize writing online. We add our contribution, which consists of:

5) Concatenate the V_4 vector to V_0 presented in the second step of the solution, in order to have finally a V_5 vector of size 23.

VI. Data Collection

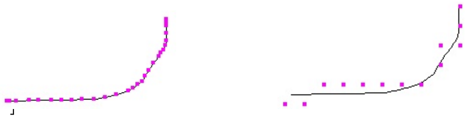
For data collection, we developed an applet (fig.7) that simulates the online writing on a tablet, and automatically generates all vectors specified in our approach based on tracking the trajectory of writing the Arabic character.



Applet démarrée.

Fig. 7. ”Da” character

The application also generates the vector point in a fixed number of resolution (See fig.8.a for resolution 1 and fig.8.b resolution 20).



(a) nbResolution=1 (b) nbResolution=20

Fig. 8. Resolution number

VII. SVM CLASSIFIER

Several techniques such as k-Nearest Neighbor (k-NN), Bayes Classifier, Neural Networks (NN), Hidden Markov Models (HMM), Support Vector Machines (SVM), etc exist for the purpose of classification. One of the commonly used techniques is SVM. Its a discriminative model that minimizes the learning error by maximizing the margin between the data classes.

VIII. EXPRIMENTATION AND TABLES RESULTS

We implemented all aspects of representation described in the previous sections. The diagram of the whole system is shown in Fig.9.

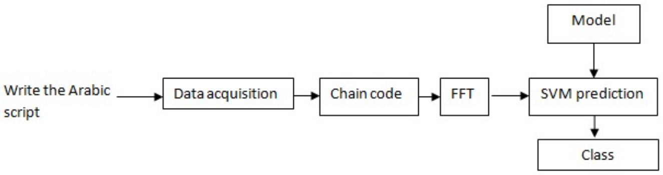


Fig. 9. Functional diagram of the recognition system showing the various

The proposed method is implemented using Java language and Statistical Pattern Recognition Tool weka. Table 1 presents the results obtained testing our OACR system on Arabic handwritten Character database. Overall recognition rate of 92.43% is achieved.

TABLE I
RECOGNITION RATES USING THE SVM CLASSIFIER

Data set	Features used	Resolution	Classifier	Recognition rate (%)
2000	FD	2	SVM	80.6%
2000	Chain codes	2	SVM	87.21%
2000	FD and Chain codes	2	SVM	92.43%

IX. CONCLUSION

The aim of this paper was to develop a new representation of shape and to use it in handwritten online Arabic character recognition. We presented a method of features extraction based on a sequence of vectors of a number of features per point to be presented to a classifier. After preprocessing the x and y coordinates on the online signal, some of local and global features are extracted. Recognition was carried out with SVM classifier. Experimental results show the high performance of our proposed scheme and the pertinence of the representation.

REFERENCES

[1] S. Madhvanath, G. Kim and V. Govindaraju,*Chaincode contour processing for handwritten word recognition*,IEEE Trans,Pattern Anal. Mach. Intell. Vol. 21, pp.928-932, 1999.

[2] I. Guyon, P. Albercht, Y. Le Cun, J. Denker and W. Hubbard, *Design of a neural network character recognizer for a touch terminal,* Pattern Recognition, volume 24, issue 2, page 105-109, 1991.

[3] M. Schenkel, I. Guyon and D. Henderson, *On-line cursive script recognition using Time Delay Neural Networks and Hidden Markov Models*,Machine Vision and Applications, special issue on Cursive Script Recognition, volume 8, pages 215-223, 1995.

[4] F. Alleau, E. Poisson, C. Viard-Gaudin and P. Le Callet, *TDNN with Masked Input*,Proc. Of the 4th Internationl Conference on Information, Communications Signal Processing and IEEE Pacific-Rim Conference on Multimedia (ICICS-PCM 2003), Nanyang Technological University, Singapour, December 2003.

[5] M. Parizeua, A. Lemieux and C. Gangné,Character Recognition Expriments Using Unipen Data,ICDAR,2001.

[6] S. Jaeger, S. Manke, J. Reichert and A. Waibel,*On-Line Handwriting Recogition : The NPen++ Recognizer*,In Internationnal Journal on Document Analysis and Recognition (IJDAR00), volume 3, pages 169-180, 2000.

[7] S. Gunter and H. Bunke,*Fast Feature Selection in an HMM-based Multiple Classifier System for Handwriting Recognition*, Pattern Recognition, Proceedings of the 25th DAGM Symposium, LNCS 2781, Springer, pages 289-296, 2003.

[8] J.J. Lee, J. Kim and J.H. Kim,*Data-driven design of HMM topology for on-line handwriting recognition*, Hidden Markov Models: Applications in Computer Vision, Volume 45, pages 107-121,2001.

[9] T. Artieres and P. Gallinari, *Stroke Level HMMs for On-Line Handwritting Recognition*,International Workshop on Frontiers in Handwritting Recognition, pages 227, 2002.

[10] S.Marukatat, *Une approche générique pour la reconnaissance de signaux écrits en ligne*, Thèse de doctorat de lUniversité Paris 6, 2004.

[11] F. Alimoglu, E. Alpaydin,*Combining Multiple Representations and Classifiers for Pen-based HandWritten Digit Recognition*,In proceedings of International Conference on Document Analysis and Recognition, pages 637-660, Ulm, Aoôt 1997.

[12] K. Badie and M. Shimura,*Machine Recognition of Arabic Hand-printed Scripts*,Trans. of IECE, Institute of Electronics and Commun. Eng. of Japan, E65, No.2, 107-114,Feb 1982.

[13] G. G. Rajput and S. M. Mali, *Marathi Handwritten Numeral Recognition using Fourier Descriptors and Normalized Chain Code*, IJCA Special Issue on ”Recent Trends in Image Processing and Pattern Recognition” RTIPPR, 2010.

[14] N. Mezghani, A. Mitiche and M. Cheriet, *Filtrage, élagage et combinaison de mémoires de kohonen pour la reconnaissance en-ligne de caractères arabes manuscrits*,In Conférence Internationale sur le Traitement et Analyse d’Images : Méthodes et Applications, TAIMA, pages 99-104, Hammamet,2003.

[15] F. P. Kuhl and C. R. Giardina, *Elliptic fourier features of a closed contour*, Computer Vision, Graphics, and Image Processing, pages 236258, 1982.

[16] H. Freeman,*Computer Processing of Line Drawings*, Computing Surveys, Vol.6, 57-97, 1990.