On-line Arabic Handwriting Recognition System based on HMM

HanyAhmed, Sherif Abdel Azeem
Electronics Engineering Department
American University in Cairo (AUC)
Cairo, Egypt
hanyahmed@aucegypt.edu, shazeem@aucegypt.edu

***Abstract*— The purpose of this research is to improve the recognition rate of on-line Arabic handwriting recognition using HMM (Hidden Markov Model). Delayed strokes are removed from the on-line Arabic word to avoid the difficulty and the confusion caused by the delayed strokes in the recognition process. Dictionaries for all the words in the ADAB database have been constructed with and without the delayed strokes. Word matching in both dictionaries along with effective on-line features and careful choice of the HMM parameters have significantly improved the recognition rate of the proposed system over other HMM-based on-line Arabic handwriting recognition systems.**

Keywords- Online handwriting recognition; Arabic; HMM

I. INTRODUCTION

In Online handwriting system, the input is a sequence of (x, y) points representing the pin-tip position while writing on a digital instrument. In recent years, some research has been done on the problem of on-line Arabic handwriting recognition [4, 5, 7]. The problem is still very challenging because of the varying writing style from person to person, difficulty of segmentation because of the cursive nature of the Arabic writing, the changing of the shape of the letter in four positions (isolated, start, middle, end), and the problem of diacritical points and marks known as delayed strokes which are usually drawn last in a handwritten word.

Due to the similarity between online handwriting and speech, many researchers have successfully applied speech recognition techniques to the handwriting recognition problem. HMM is the most popular technique for speech recognition used recently for handwriting recognition [2]. HMM avoids the problem of pre-segmentation of words into characters in both the training and the classification stages, so the errors of pre-segmentation can be eliminated.

Our target is to study the recognition of online Handwritten Arabic words of the ADAB-database [1] using HMM theory. The ADAB-database in version 1.0 is spilt into 3 sets. Details about the number of files, words, characters, and writers for each set 1 to 3 are shown in Table I.

TABLE I. FEATURES OF ADAB-DATASETS 1, 2, AND 3

Set	Files	Words	Characters	Writers
1	5037	7670	40500	56
2	5090	7851	41515	37
3	5031	7730	40544	39
Sum	15158	23251	122559	132

II. PRE-PROCESSING

The pre-processing step is intended to reduce noise, solve the problem of various speeds of writing and increase the ease of extracting features. The pre-processing step involves interpolating the missed points, smoothing, baseline extraction, and removing the delayed strokes as shown in Fig.1

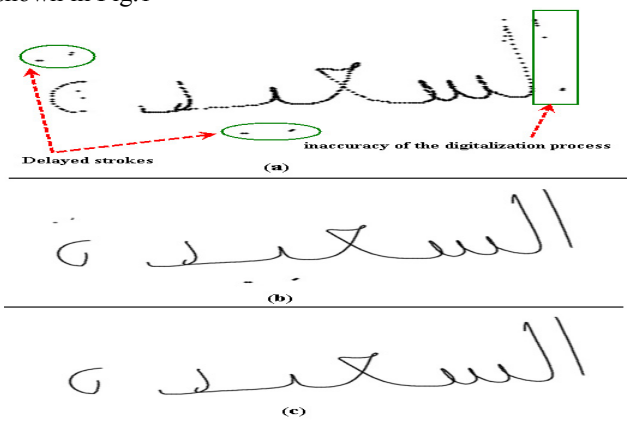

Figure 1. Example of solving the problem of inaccuracy of the digitalization process and removing the delayed strokes : (a) Original word, (b) Word after interpolating the missed points,(c) Word after removing the delayed strokes

A. *Interpolation* [8]:

Linear interpolation is used to solve the problem of inaccuracy of the digitization process and writing speed normalization. Interpolation is done for each stroke separately, as shown in Fig.1.b.

B. *Smoothing :*

Smoothing is the second step of pre-processing and it is used to eliminate noise. A 5-point moving average algorithm is used for this purpose. The moving average filter smoothes data by replacing each data point with the average of the neighboring data points defined within the span of the filter. This process is equivalent to low-pass filtering with the response of the smoothing given by the difference equation

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.266

IEEE computer society

$$y(i)=\frac{1}{2N+1}(y(i+N)+y(i+N-1)+...+y(i-N))$$

Where y (i) is the smoothed value for the ith data point, N is the number of neighboring data points on either side of y (i), and 2N+1 is the span.

C. Baseline extraction [9]:

X-Y points of the given word were converted into its corresponding bitmap matrix, and then horizontal projection algorithm has been followed, as shown in Fig. 2.

Figure 2. Extracting the baseline using horizontal projection.

D. Removing the delayed strokes:

Delayed strokes are a well-known problem in online handwriting recognition [4, 6]. Those strokes complicate online recognition because the writing order of the delayed strokes is not fixed and varies among different writers. Figure 3 illustrates the delayed strokes found in the ADAB database.

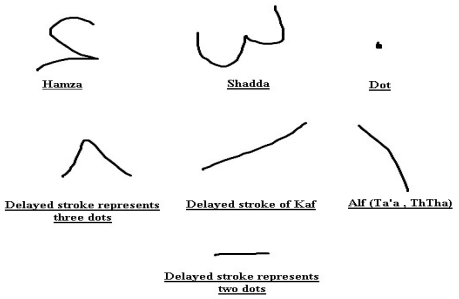

Figure 3. The delayed stroke of the ADAB database

Delayed strokes are detected by comparing the height, width, bounding box area, position from the baseline, and number of points for each stroke after pen up position with pre-determined thresholds as shown in Fig.4.

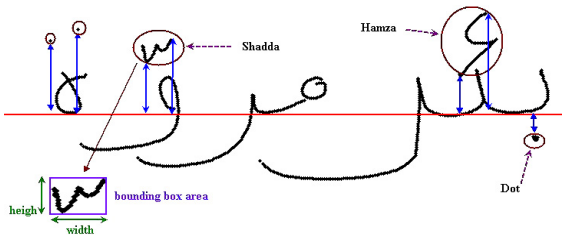

Figure 4. Example of rules used to extract the delayed strokes.

III. HMM

Our system is based on HMM implemented using the HMM Toolkit (HTK Engine). We used left-to-right HMMs for modeling the Arabic characters. Arabic contains 28 letters, written horizontally from right to left. Most Arabic letters have four different shapes, depending on their position within a word which results in more than 100 HMM models. After removing the delayed strokes, the number of models decreased to 64 models (as described in Table II) because similar characters were grouped together in one class such as beginning Ba'a, Nun, Ta'a, and Tha'a as shown in Fig.5

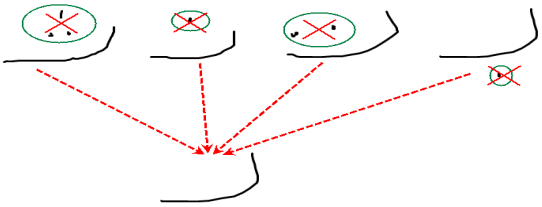

Figure 5. Grouing the characters in classes after removing the delayed strokes

TABLE II. ARABIC CLASSES

Letter and ligatures	Possible Shapes			
	Isolated	*End*	*Middle*	*Start*
Aleph	ا أ إ آ	ـا ـأ		
Ba'a ,Ta'a , Tha'a,Nun,Ya'a	ب ت ث ن ي	ـب ـت ـث ـن ـي	ـبـ ـتـ ـثـ ـنـ ـيـ	بـ تـ ثـ يـ نـ
Jeem,Ha'a,Kha'a	ج ح خ	ـج ـح ـخ	ـجـ ـحـ ـخـ	جـ حـ خـ
Dal , Thal	د ذ	ـد ـذ		
Raa , Zai	ر ز	ـر ـز		
Seen ,Sheen	س ش	ـس ـش	ـسـ ـشـ	سـ شـ
Sad , Dad	ص ض	ـص ـض	ـصـ ـضـ	صـ ضـ
TTa,ThTha	ط ظ	ـط ـظ	ـطـ ـظـ	طـ ظـ
Ein , Gein	ع غ	ـع ـغ	ـعـ ـغـ	عـ غـ
Faa ,Qaf	ف	ـف	ـفـ ـقـ	فـ قـ
Qaf	ق	ـق		
Kaf	ك	ـك	ـكـ	كـ
Lam	ل	ـل	ـلـ	لـ
Meem	م	ـم	ـمـ	مـ
Nun	ن	ـن		
Hah , Ta'a	ه ة	ـه ـة	ـهـ	هـ
Waw	و	ـو		
Ya'a	ى ي	ـى ـي		
Hamza	ء			
Lam Alf	لا لإ لأ لآ	ـلا ـلإ ـلأ ـلآ		
Meem Ha'a				
Lam Ha'a				
Lam Meem				

As shown in Table 2; isolated characters have 20 classes, beginning characters have 14 classes, middle characters have 11 classes, and end characters have 19

classes. Besides the previous classes, the ADAB database has some digits and Latin letter "V".

Figure 6 displays the case of a 20 states HMM for each character used in our system.

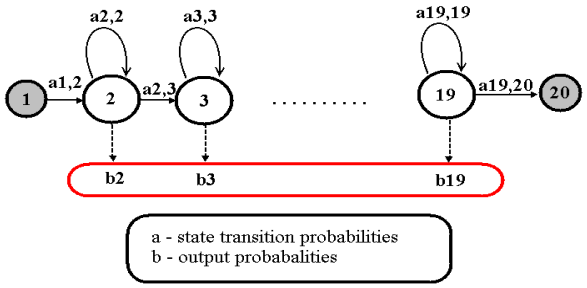

Figure 6. Left-to-right states

18 states of these are emitting states and have output probability distributions associated with them. HTK is principally concerned with continuous density models in which each observation probability distribution is represented by a mixture Gaussian density. In this case, for state j the probability $b_j(o_t)$ of generating observation o_t is given by:

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} \Psi(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s}$$

Where M_{js} is the number of mixture components in state j for stream s, The exponent γ_s is a stream weight and its default value is one, c_{jsm} is the weight of the m□th component and ψ (o; μ, □) is a multivariate Gaussian with mean vector μ and covariance matrix □, that is

$$\Psi(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1} (o-\mu)}$$

Where n is the dimensionality of o.

16-Mixtures have been chosen after large number of experiments to give a robust model for each character and high recognition rate. The transition matrix for this model has 20 rows and 20 columns. Each row will sum to one except for the final row which is always all zero since no transitions are allowed out of the final state.

IV. System Description

E. Dictionary of words with their delayed strokes (dict1)

A dictionary of all the different unique words in the ADAB database along with their delayed strokes has been constructed. Figure 7 shows a flow chart that describes the sequence of constructing the dictionary of the available words in the ADAB database with their delayed strokes.

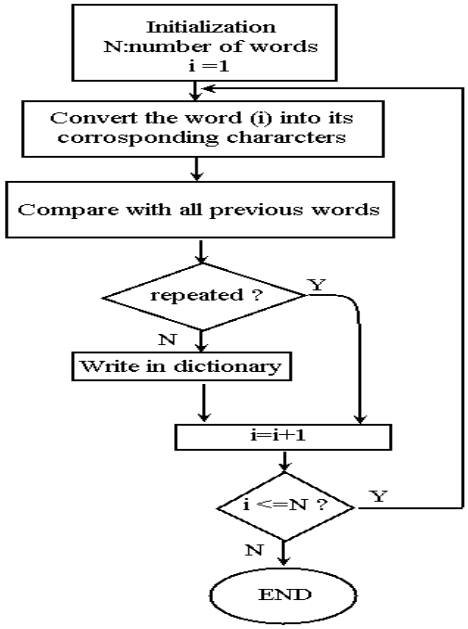

Figure 7. Flow chart of the algorithm used to construct the dictionary of the unique set of words in the ADAB database with their delayed strokes.

The previous algorithm results in a dictionary (dict1) with unique set of words with their delayed strokes as shown in Fig.8.

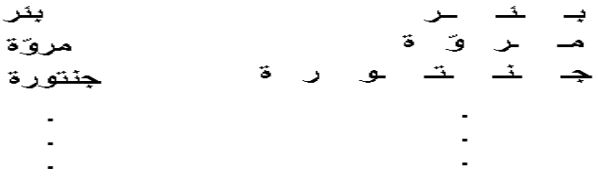

Figure 8. Dictionary of words with the delayed strokes (dict1).

F. Dictionary of words without delayed strokes (dict2)

Another dictionary containing all the different words in the ADAB database without their delayed strokes has been constructed for training and testing purposes. Figure 9 illustrates the flow chart that describes the sequence of constructing the dictionary of words without delayed strokes from the previous dictionary (dict1).

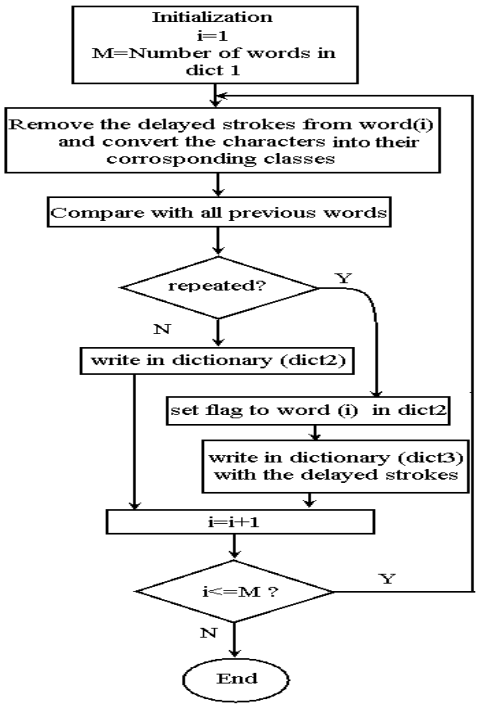

Figure 9. Flow chart of the algorithm used to construct the dictionary of unique set of words in the ADAB database without their delayed strokes.

The previous algorithm results in a dictionary (dict2) which contains a unique set of words without their delayed strokes, as shown in Fig. 10. While constructing the dictionary (dict2), it was discovered that some words coming from dict1 become exactly the same after removing their delayed strokes. For example, the words "متلين, مثلين " become identical when they have their delayed strokes removed. Hence, a third dictionary (dict3) has been constructed for those words only and a flag was added to all the words in dict2 and the flag was set only for those words that confuse with other words after the removal of their delayed strokes as illustrated in the flow chart shown in Fig. 9. Figure 11 shows some of those words in dict3.

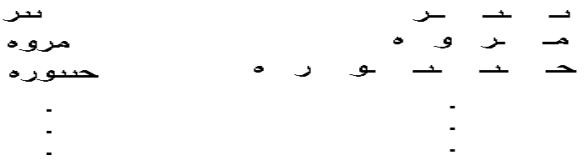

Figure 10. Dictionary of words without the delayed strokes (dict2).

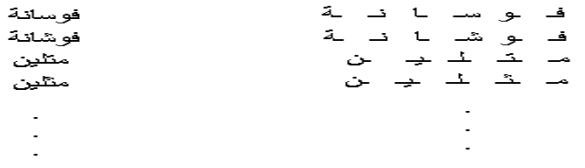

Figure 11. Dictionary of words which confuse with other words after removing their delayed stroks (dict3).

G. *Classification stages*

The classification process goes through two stages. The first stage includes removing the delayed strokes from the input test word, extracting its on-line features, and then classifying the word by choosing the best candidate from dict2. If the flag of the best candidate is set, the system will go through the second stage. The second stage searches among the words in dict3 corresponding to the top candidate in dict2. The search among the available words in dict3 is based on comparing the delayed strokes of the input test word with the delayed strokes of dict3 candidates. Comparing the delayed strokes depends on three rules; the number of the delayed strokes, their position, and their shape (Dot, line, Hamza …etc). The classification stages are illustrated in the block diagram given in Figure 12.

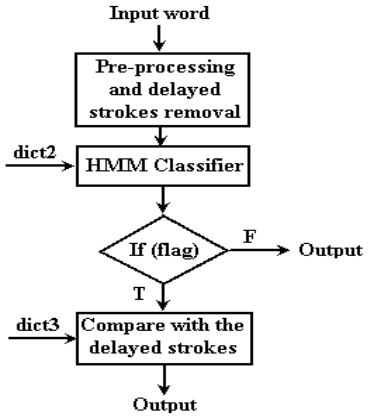

Figure 12. Block diagram shows the two stages of the classification.

V. FEATURE EXTRACTION

The following steps describe our method to extract the feature vector. Let $P = (X_i, Y_i)$, i=1, 2, 3 …N, where N is the number of points that represent the main body of the input word without delayed strokes. The local writing direction at a point at instant t is described by [6]:

$$\cos(\alpha(t)) = \frac{\Delta Y\ (t)}{\Delta S\ (t)}$$

$$\sin(\alpha(t)) = \frac{\Delta X\ (t)}{\Delta S\ (t)}$$

$\Delta X\ (t)$, $\Delta Y\ (t)$ and $\Delta S\ (t)$ are defined as follows:

$$\Delta X\ (t) = X\ (t-1) - X\ (t+1)$$

$$\Delta Y\ (t) = Y\ (t-1) - Y\ (t+1)$$

$$\Delta S\ (t) = \sqrt{\Delta X^{\,2}(t) + \Delta Y^{\,2}(t)}$$

Delta and acceleration coefficients [2] were calculated and then concatenated with the directional features.

VI. RESULTS

In our experiments, we used two sets of the ADAB database for training and the third one for testing. Table III reports the used sets and the corresponding recognition accuracies.

TABLE III. RECOGNITION RATE OF THE PROPOSED SYSTEM

Training Sets	Test Set	Accuracy
1,2	3	95.27%
1,3	2	89.72%
2,3	1	92.71%

The results obtained by the proposed system clearly outperform those obtained by other HTK-based recognition systems as shown in Table IV. REGIM-HTK and REGIM-CV-HTK are two HMM-based systems that used HTK as a recognizer. Both systems participated in the ICDAR 2009 Online Arabic Handwriting Recognition Competition [1]. Table IV reports the recognition results (top 1) achieved by both systems on the ADAB database.

Table IV. Recognition rates of other HTK-based systems

System	Set1	Set2	Set3
REGIM-HTK [4,5]	57.87%	54.26%	53.75%
REGIM-CV-HTK [3,7]	28.85%	35.75%	30.6%

The misclassifications obtained by the proposed system are due to two main reasons: the removal of the delayed strokes and the lack of training data for some of the classes in the ADAB database. Removing the delayed strokes results in misclassification because some small characters are confused as delayed strokes and vice versa. Moreover, some writers in the ADAB database move their hands up while writing one character, so sub-strokes may be detected as delayed stroke, as shown in Fig. 13.

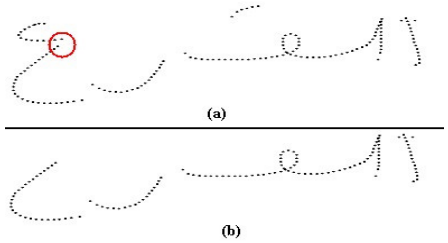

Figure 13. Problem of moving the hand up while writing one character: (a) Original word , (b) word after removing the delayed strokes and the sub-stroke of the isolated Ein.

The other problem is the insufficiency of the training some classes in the database such as isolated (Sad, Dad), isolated (Ein, Gein), end (Ha'a, Kha'a, Gem), end (Sad, Dad), and end (TTa, ThTha) which leads to poor test results.

VII. CONCLUSION

This paper presents an HMM-based on-line Arabic handwriting recognition system. The main feature of the proposed system is the removal of the delayed strokes in the training and test phases in order to avoid the confusion caused by the vast differences in the order of the writing of the delayed strokes by different writers. Careful choice of the HMM parameters and efficient on-line directional, delta, and acceleration features have significantly improved the performance of the proposed system over other HMM-based systems.

REFERENCES

[1] Haikal El Abed, Volker Margner, Monji Kherallah, and Adel M. Alimi , "ICDAR 2009 Online Arabic handwriting recognition competition," In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition (ICDAR '09).

[2] Nick Bardici, and Björn Skarin, " Speech Recognition using Hidden Markov Model," M.Sc.Thesis, Department of Telecommunications and Signal Processing , Blekinge Institute of Technology, 2006.

[3] Houcine Boubaker , Monji Kherallah, and Adel M. Alimi , "New algorithm of straight or curved baseline detection for short arabic handwritten writing," In Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), pp. 778–782, July 2009.

[4] Abdelkarim Elbaati, Houcine Boubaker, Monji Kherallah, Adel M. Alimi, Abdellatif Ennaji, and Haikal El Abed ,"Arabic handwriting recognition using restored stroke chronology," In Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), pp. 411–415, July 2009.

[5] Mahdi Hamdani, Haikal El Abed, Monji Kherallah, and Adel M. Alimi, " Combining multiple HMMs using online and offline features for offline Arabic handwriting recognition," In Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 201–205, July 2009.

[6] S. Jager, and S. Manke, J. Reichert, and A. Waibel, "Online handwriting recognition: the npen++ recognizer," IJDAR, vol. 3, no. 3, pp. 169–180, 2001.

[7] Monji Kherallah, F. Bouri, and Adel M. Alimi, " On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm," Eng Appl. Artif. Intell. 22, 153–170 (2009).

[8] Moisés Pastor, Alejandro Toselli, and Enrique Vidal, "Writing speed normalization for on-Line handwritten text recognition," Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05).

[9] Samia Snoussi Maddouri,Fadoua Bouafif Samoud,Kaouthar Bouriel,Noureddine Ellouze, and Haikal El Abed, "Baseline Extraction: Comparison of six methods on IFN/ENIT Database," The 11th International Conference on Frontiers in Handwriting Recognition,2008.