Segmentation-Free Online Arabic Handwriting Recognition

FADI BIADSY

Computer Science, Columbia University
New York,NY 10027,USA.
fadi@cs.columbia.edu

RAID SAABNI

Computer Science, Ben-Gurion Unviversity of The Negev,Beer-Sheva,84105,Israel.
Triangle Research&Development Center, Kafr Qara,30075, Israel.
saabni@cs.bgu.ac.il

JIHAD EL-SANA

Computer Science, Ben-Gurion Unviversity of The Negev
Beer-Sheva,84105,Israel.
el-sana@cs.bgu.ac.il

Arabic script is naturally cursive and unconstrained and, as a result, an automatic recognition of its handwriting is a challenging problem. The analysis of Arabic script is further complicated in comparison to Latin script due to obligatory dots/stokes that are placed above or below most letters. In this paper, we introduce a new approach that performs online Arabic word recognition on a continuous word-part level, while performing training on the letter level. In addition, we appropriately handle delayed strokes by first detecting them and then integrating them into the word-part body. Our current implementation is based on Hidden Markov Models (HMM) and correctly handles most of the Arabic script recognition difficulties. We have tested our implementation using various dictionaries and multiple writers and have achieved encouraging results for both writer-dependent and writer-independent recognition.

Keywords: Online Handwriting Recognition; Arabic; HMM.

1. Introduction

Keyboards and electronic mice may not endure as the prevalent means of human-computer interfacing. Devices such as digital tablets, hand-held computers, and mobile technology, provide significant opportunities for alternative interfaces that work in forms smaller than the traditional keyboard and mouse. In addition, the need for more natural human-computer interfaces becomes ever more important as computer use reaches a larger number of people. Two such natural alternatives to typing are speech and handwriting, which are universal human communication methods. Both are potentially easier human-computer interfaces to learn by new users compared to keyboards. Although a handwriting interface expects users to be literate, it ensures a higher degree of privacy and confidentiality compared to

speech.

Automatic handwriting recognition has been classified into two categories, *offline* and *online*, based on the presentation of the data to the system. *Offline handwriting recognition* approaches do not require immediate interaction with users. A scanned handwritten or printed text is fed to the system in a digital image format. In a typical *online handwriting recognition* approach, a special stylus is used to write on a digital device, such as a digital tablet. The digitized samples are fed to the system as a sequence of 2D-points in real-time, thus tracking additional temporal data not present in offline input.

In this paper we extend the work [12,13] and introduce an online handwriting recognition system for Arabic script, which is used in various languages, such as Arabic, Farsi, Urdu, Pashto, and Kurdish. Our approach performs the recognition on the continuous word-part level and the training on the letter level. Such a scheme avoids the segmentation of words into individual letters during the recognition process, which is often prone to errors, and substitutes the training for large set (the word-parts) by a small set (the letters). Figure 1 depicts the flow of our recognizer. Our approach accurately handles delayed strokes by first detecting them and then integrating them into the word-part body. The current implementation is based on Hidden Markov Models (HMM) and deals with many of the Arabic script recognition difficulties. We focus on word-level recognition of undiacritized (unvocalized) Arabic, and thus no sentence-level context is modeled. Arabic vocalic diacritics are most often ignored in writing and printing and, therefore, not addressed here. In this work we treat the *shadda* (ω) as diacritic and thus it is not addressed in this work.

In the rest of the paper, we first explain the basic characteristics of the Arabic script followed by an overview of related work in handwriting recognition. Then, we discuss preprocessing and feature extraction, the recognition framework, and evaluation results. Finally, we draw some conclusions and suggest directions for future work.

2. Characteristics of the Arabic Script

Arabic script consists of 28 basic letters, 12 additional special letters, and 8 diacritics[a]. Arabic is written (machine printed and handwritten) in a cursive style from right to left. Most letters are written in four different letter shapes depending on their position in a word, e.g., the letter ع (Ain)[b] appears as ع (isolated), عـ (initial), ـعـ (medial), and ـع (final). Among the basic letters, six are Disconnective – ا (Alef), د (Dal), ذ (Thal), ر (Reh), ز (Zain), and و (Waw). Disconnective letters do not connect to the following letter and have only two letter shapes each. The presence of these letters interrupts the continuity of the graphic form of a word. We denote connected letters in a word, as a *word-part*. If a word-part is composed of only one

[a]The diacritics are not explored here, since they are almost never used in handwriting.
[b]All Arabic letters are transliterated in Buckwalter's Arabic transliteration format, without diacritics (refer to www.ldc.upenn.edu/myl/morph/buckwalter.html)

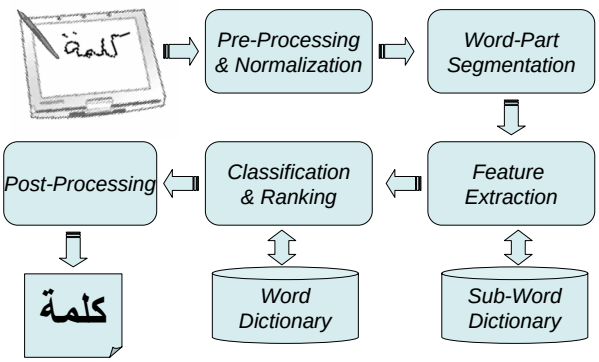

Fig. 1. The main stages of our online Arabic handwritten recognizer

letter, this letter will be in its isolated shape. For example, the Arabic word مرتفعات (mrtfEAt) "heights" consists of 7 letters (from right to left): م (Meem), ر (Reh), ت(Teh), ف(Feh), ع (Yeh), ا (Alef), and ت (Teh), which are realized initially مـ, finally ـر, initially تـ, medially ـفـ, medially ـعـ, finally ـا, and isolated ت, respectively. This word has three word-parts (from right to left): مر, تفعا, and ت.

Arabic script is similar to Roman script in that it uses spaces and punctuation marks to separate words. However, certain characteristics relating to the obligatory dots and strokes of the Arabic script distinguish it from Roman script, making the recognition of words in Arabic script more difficult than in Roman script. First, most Arabic letters contain dots in addition to the letter body, such as ش (Sheen) which consists of س (Seen) letter body and three dots above it. In addition to dots, there are strokes that can attach to a letter body creating new letters such as ك, ط, and لا. These dots and strokes are called *delayed strokes* since they are usually drawn last in a handwritten word-part/word. Second, eliminating, adding, or moving a dot or stroke could produce a completely different letter and, as a result, produce a word other than the one that was intended (see Table 1). Third, the number of possible variations of delayed stokes is greater than those in Roman script, as shown in Figure 2. There are only three such strokes used for English: the cross in the letter t, the slash in x, and the dots in i and j.

Finally, in Arabic script a top-down writing style called *vertical ligatures* is very common – letters in a word may be written above their consequent letters. In this style, the position of letters cannot be predefined relative to the baseline of the word. This further complicates the recognition task, particularly in comparison with the Roman script. Due to the holistic approach in our proposed recognition model, no restrictions were applied regarding the top-down writing style.

	1	2
a	عزام	غرام
b	عرب	غرب

Table 1. Word (a_1) (EzAm) "lion" is a result of moving the dot to the left from word (a_2) (grAm) 'love'. Word (b_1) (Erb) "Arab" is a result of eliminating the dot above the first letter from word (b_2) (grb) "west".

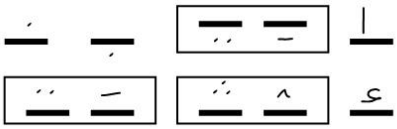

Fig. 2. Delayed strokes in Arabic script may appear under or above the letter body. The boxed pairs represent common variants (e.g., three dots are often written as a circumflex "hat"). These seven strokes appear in letters used in writing standard Arabic. Eleven additional strokes exist for writing additional letters in other languages (Urdu, Pashto, Farsi, etc.)

3. Related Work

For the last three decades Hidden Markov Models (HMM) were successfully used for automatic speech recognition. Due to the success of HMM in modeling sequential data It has been also adopted for modeling letters in handwriting recognition. Many variations of HMM models had been adapted and used in script recognition research. Discrete, continuous, and semi-continuous types were used with various topologies ranging from ergodic to left-to-right models with no state skipping. HMM-based algorithms were designed to handle letters, words, strokes or pseudo characters using one dimensional, two dimensional or planner Hidden Markov Models. Results were very encouraging in the handwritten case and appear to handle the cursiveness well.

Pechwitz and Maergner [31] presents an offline recognition system for Arabic handwritten using a semi-continuous HMM. They used a sliding windows that moves from right to left to collect features directly from the normalized gray image pixels. Then they apply Loeve-Karhunen transformation to reduce the number of features in each frame. They used seven states model for each character shape. Tests were performed using the IFN/ENIT database of handwritten Arabic words and achieved 89% maximal recognition rate. Khorshed [22] used the Hidden Markov Model Toolkit (HTK) to develop offline printed Arabic text recognition system. After decomposing the document image into text line images, a narrow sliding window is used to extract a set of simple statistical features. The system was applied to a

data corpus which includes Arabic text of more than 600 A4-size sheets typewritten in multiple computer generated fonts and achieved 95% maximal recognition rate.

Mahmoud [25] used HMM based system to recognize offline handwritten Arabic (Indian) numerals. Angle, distance, horizontal, and vertical span features were extracted from these numerals as units for training and testing the HMM. The number of states had been estimated by performing several experiments which show that the best results were achieved using an HMM model with 10 states. The system achieved an average recognition rate of 97.99% on a large private database using 120 features presented as 12 observations of 10 features per digit. Benouareth *et al.* [11] presented an offline segmentation-free recognition system for unconstrained Arabic handwritten words using discrete HMM with explicit state duration. The explicit state duration modeling were used to improve the discriminating capacity of the HMM and enable the recognition of difficult pattern in an unconstrained Arabic handwriting. They used a new version of the Viterbi algorithm that takes into account explicit state duration to perform efficient training and testing tasks. A set of statistical and structural features were extracted from the word image using a sliding window approach based on vertical projection histogram. Experiments using the IFN/ENIT database achieved 90.2% average recognition rate.

Al-Hajj *et al.* [2] presented a segmentation-free system for offline recognition of cursive Arabic handwritten words. They used three HMM-based classifiers and combination of their results are used to determine the recognized words. Sliding windows with different orientations were used to extract pixel-level features, such as pixel density distribution and local pixel configurations from the binary image. They have tested different combination schemes and have achieved a recognition rate of up to 90.96% on the IFN/ENIT database.

Khorsheed [21] presented a segmentation-free method for offline recognition of cursive handwritten Arabic script. Structural features were extracted from the skeleton of the words after segmentation to elementary strokes. These features are used to train a single hidden Markov Model. The HMM is composed of multiple character models where each model represents one letter from the alphabet. The proposed method achieved recognition rate of 72% and 87% after consulting with a word dictionary on samples extracted from a historical handwritten manuscript [20]. Al-Muhtaseb *et al.* [3] proposes a HMM-based system for offline recognition of Arabic printed texts. Sixteen features were generated from each vertical sliding strip from overlapping and non overlapping hierarchical windows. Eight different fonts were used for training and testing and yield recognition rates around 99%. Dehghan *et al.* [14] presented a segmentation-free approach for offline handwritten Farsi/Arabic words recognition. A discrete HMM was used for the recognition process and Kohonen self-organizing Maps for vector quantization of the feature vectors. A sliding window on the histogram of the chain code directions was used to generate the feature vectors. The width of the sliding window was fixed to twice the stroke width and divided to five horizontal zones. Experiments carried out on test samples of $17,000$ of 198 city names in Iran achieved 65.05% recognition rate.

Menasri *et al.* [27] present a hybrid system based on HMM and neural network classification methods using explicit grapheme segmentation. Each letter-body class is represented by an HMM model and the Neural network computes the observations probability distribution. Experiments using IFN/ENIT database result in 87% average recognition rate. Dots and diacritics were recognized independently and used as prior knowledge to eliminate and validate letters.

Some papers focused on the recognition of isolated forms of Arabic letters or Digits only [8,16,29,28,5,9,10]. Recently, many attempts develop algorithms to recognize the cursive form of the Arabic script. HMM-based system received most of the attention, but other techniques were also used and proved to have satisfying results. Al-Emami and Usher [1] developed an online Arabic handwriting recognition system, based on decision-tree techniques. Their system was tested on 13 Arabic letter shapes. Alimi [6] developed an online writer-dependent system to recognize Arabic cursive words using a neuro-fuzzy approach. The system was tested using one writer on 100 replications of a single word. Al-Taani [4] used a structural approach to develop an online Arabic digit recognizer. Primitives representing specific strings are extracted from each digit. Then, grammars, constructing these strings, are used to identify digits. The system was tested using 100 different writers, and an average recognition rate of 95% was reported. Mezghani *et al.* [28] developed an online recognition system for isolated Arabic letters using Fourier descriptors and Kohonen maps. They reported a recognition rate of 86% on 7244 samples of 17 classes written by 17 writers. Fourier descriptors and tangents, extracted along the boundary, were used to represent the characters. Alimi and Ghorbel [5] developed an online recognition system for isolated Arabic characters using dynamic programming algorithms. They reported a recognition rate of 93% using different database sizes and replication of characters.

AraPen is an Arabic online handwriting recognition system, which was developed by Alsalakh and Safadi [7]. The system, which is based on Dynamic Time Warping (DTW), was designed to handle non-cursive character recognition and adapted to the cursive case. In the non-cursive case, the system was tested on a small corpus and achieved a recognition rate of 91%, after training with a specific writer's style. The recognition rates went down dramatically, to lower than 50%, when adapting the system to cursive scripts. Baghshah *et al.*[9] developed a system to recognize isolated Persian handwritten letters online. The system is based on a fuzzy logic approach and yields a recognition rate of 95% using the Razavi and Kabir database [33]. Halavati *et al.* [17] used visual features and fuzzy logic classifiers to develop a system for online recognition of Persian handwriting.

In general, previous work has viewed delayed strokes as features that add complexity to online handwriting recognition. Four methods were proposed to recognize words with delayed strokes:

- Delayed strokes were totally discarded from handwriting in the preprocessing phase [6].

- Delayed strokes were detected in the preprocessing phase and then used in a post-processing phase [19].
- The end of a word was connected to the delayed strokes with a special connecting stroke [26]. Adding the special stroke, which indicates that the pen was raised, results in a continuous-stroke sequence for the entire handwritten English sentence.
- Delayed strokes were treated as special characters in the alphabet [19], i.e., a word with delayed strokes was given alternative spellings to accommodate different sequences where delayed strokes are drawn in different orders.

These four methods are not adequate for recognizing Arabic script. The first and second methods cannot be employed effectively since the number and location of dots define the letter itself. Eliminating delayed strokes causes tremendous ambiguity, particularly when the letter body is not written clearly. Furthermore, eliminating delayed strokes may lead to a similar shape that may represent different letters or a sequence of letters. For example, the letter (Seen) سـ has a shape similar to that of the three letters بتيـ (b + t + y) (without dots) in some writing styles. The third and fourth methods also cannot be implemented, since Arabic words may contain many delayed strokes. These methods dramatically increase the hypothesis space, since words should be represented in all their handwriting permutations. For example, the word حقيقية (Hqyqyp) 'truth' contains 10 dots, 6 are above the word and 4 under it. Connecting the delayed strokes with the end of the word complicates the representation of the word and removing the delayed strokes (dots) require handling $3 \times 4 \times 5 \times 4 \times 2 = 480$ different representation.

4. Preprocessing and Feature Extraction

In this section, we describe our geometric preprocessing, feature extraction, and our novel solution for delayed-strokes.

4.1. *Geometric Preprocessing*

Acquired point sequences pass a geometric processing phase to minimize handwriting variations. We have used a low-pass filter algorithm [34] to reduce noise and remove imperfections caused by acquisition devices. To simplify the point sequences, in order to eliminate redundant points irrelevant for pattern classification,

we applied the Douglas and Peucker algorithm [15]. To complete the preprocessing, we performed writing-speed normalization by re-sampling the point sequences.

4.2. *Feature Extraction*

We extract three main features for each point in the processed point sequence: *local_angle*, *super_segment*, and *loop_presence*. The *local_angle* feature of p_i, which is denoted by $local\_angle_i$, is defined as the angle between the segment $\overline{p_{i-1}, p_i}$ and the x-axis ($i > 1$), as shown in Figure 3. The *super_segment* feature provides wider geometric information which relates each segment to its segment group. This feature is computed by further simplifying the processed point sequence to obtain a coarse representation, which will be donated the *skeleton points*. Every two consecutive skeleton points define a skeleton segment. The super-segment feature for point p_i, which temporally appears between two consecutive skeleton points, is defined as the angle between the skeleton segment and the x-axis (see Figure 3). This feature is denoted by $super\_seg\_angle_i$. The *loop feature* is a global feature that indicates the presence of a loop in the processed point sequence. Global features capture information related to the global geometric shape of the whole word/letter. Three common global features have been used in previous work for handwriting recognition: loops, cusps, and crossings [18]. In this work, only the loop feature is used, since loops are obligatory in many Arabic letter shapes, e.g., (f) ف. In contrast, cusps and crossings are less common and vary among writers. Global features are not robust features by themselves for unconstrained script. However, the loop feature has greatly improved our recognition rate. We will refer to this feature for point p_i as is_loop_i, which is 1 if p_i is in a loop, otherwise 0. Different people write the initial form of the Arabic Letters ح , خ and ج while performing a loop while others writes them similar to the printed form with no loops. In this work we restrict the writing style to the case not performing these loops. Extending this work to include such cases can be done by treating the different shapes as two different letters (with and without loop) representing the same letter for each one of these three letters.

4.3. *Handling Delayed Strokes*

Delayed strokes are essential to distinguish between various Arabic letters. Thus, handling them correctly is vital for accurate recognition of the Arabic script. We have developed the *delayed-stroke projection* algorithm to integrate a delayed stroke within the appropriate word-part body. Our algorithm involves two steps, the detection of delayed strokes and the incorporation of delayed strokes within the right word-part body of the processed point sequence.

In Arabic scripts, delayed strokes are written above or below a word-part and could appear before, after, or within a word-part with respect to the horizontal axis, as shown in Figure 2. Usually, delayed strokes are written immediately after completing the word-part body. This creates the general interleaved sequence

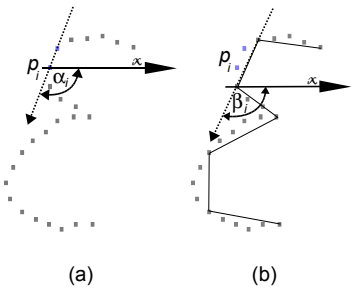

Fig. 3. The $local\_angle_i$ and $super\_segment_i$ features, denoted by α_i and β_i, repectively .

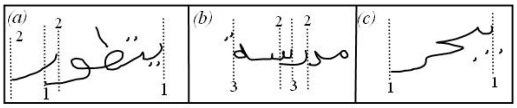

Fig. 4. Possible delayed stroke positions used for detection: (a) five delayed strokes for word-part 1; (b) two delayed strokes for word-part 3; (c) three delayed strokes for word-part 1.

$wp_1, ds_1, wp_2, ds_2, \cdots, wp_n, ds_n$ where wp_i is $i-th$ word-part and ds_i is the $i-th$ delayed-stroke set associated with wp_i. The delayed-stroke set can be empty for word-parts without delayed strokes. To detect the delayed strokes associated with a word-part, it is enough to determine whether a given processed point sequence forms a delayed stroke or not.

The detection of delayed strokes – dots and short-strokes – is performed based on their sequential order, location, and size. Dots are detected based on the size and shape of their bounding box with respect to the word-part. They usually tend to have nearly square bounding boxes. Valid non-dot delayed strokes are required to either fall within the horizontal boundary of the word-part or to appear before (on the right side of) the word-part. This restriction allows consecutive word-part bodies to overlap, as shown in Figure 4 (a) and (b), e.g., in (a) word-parts 1 and 2 overlap.

Upon the detection of a delayed stroke and distinguishing it from the word-part body, we perform the *delayed-stroke projection*, which is illustrated in Figure 5 (with one letter). Our delayed-stroke projection algorithm starts by vertically projecting the first point of the delayed stroke q_1 into the letter body at point p_i. Then incorporating the delayed stroke into the letter body by inserting the delayed-stroke's point sequence, d_{ps}, into the letter-body's point sequence, l_{ps}, starting from p_i. Finally, the last point of the delayed stroke is connected to point p_{i+1}. The two newly added virtual segments that connect the delayed stroke with the letter body are uniformly sampled (according to the average sampling rate of the

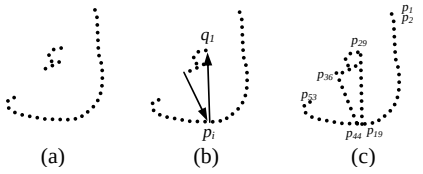

Fig. 5. (a) The projection of the delayed stroke ء in the letter ك (k); (b) the delayed stroke is projected to the letter body; (c) the newly generated PPS (p1 to p53).

Fig. 6. Three delayed strokes are projected in the second and third word-part bodies for the handwritten word: الانطباع (AlAnTbAE) 'the impression'.

word-part). We will denote the points on the virtual segment as the *virtual points*.

Arabic letters usually appear within connected word-parts and not as isolated letters. Delayed-stroke projection is used to integrate a delayed stroke within a word-part body, as in the isolated-letter case (see Figure 6). In the cases where a delayed stroke appears before or after the word-part body, as shown in Figure 4 (b) and (c), we connect the delayed stroke to the closest point of the word-part body.

4.4. *Feature-Vector Construction*

Due to the nature of our feature space, we have adopted the discrete Hidden Markov Model[c] (HMM) (for a tutorial in Hidden Markov Model, please refer to [32]) for the recognition task. The input to this model is a sequence of discrete values, which are usually denoted as the observation sequence. Thus, a quantization process is required to convert the three-dimensional feature-vector sequence, extracted from a handwritten word-part, to a discrete observation sequence. In our current implementation, each observation o_i in an observation sequence is an integer value $[0 \cdots 259]$(which are represented using 9 *bits*). This sharp discretization is necessary to reduce the training samples for online Arabic handwriting systems. The lowest 8 *bits* are used to represent the 3D-feature vector – $local\_angle_i$, $super\_seg\_angle_i$, and is_loop_i. The $local\_angle_i$ and $super\_seg\_angle_i$, which are real angle values, are converted to 16 and 8 directions, respectively (similar to [23]); and the feature is_loop_i is a binary value (one bit). The $9th$ *bit* is used to mark virtual points and when it is on, the first two bits are used to describe the property of the corresponding virtual point. Th observation values $[256 \cdots 259]$ (which correspond to the first two bits $00 \cdots 11$, respectively) are used to classify the virtual points using (a) the position of the delayed stroke (above or below the word-part), and (b) the direction

[c]Using a multi-variant mixture of Gussians (as emission probability density function) will not model properly the binary feature, the is_loop in our case

D_1 = {هل , معلم , محمد}
D_2 = {محمود , جامعة , ثقافة}
D_3 = {وسام ,فادي , التحدي , انسان}
D_4 = {رواية}

Fig. 7. The subdivision of a dictionary into sub-dictionaries.

of the virtual segment (up or down). These four observation values are crucial to distinguish between different letter shapes that have the same letter body, but differ on the position of their delayed strokes, e.g., تـ(Teh) and يـ(Yeh).

5. Recognition Framework

Our recognition framework uses discrete HMMs to represent each letter shape. To enhance word recognition, these letter-shape models are embedded in a network that represents a word-part dictionary. The recognition of word-parts is performed without explicit segmentation into letter shapes, and instead, the recognition is performed along paths that represent valid word-parts, similar to [18,26,30]. Our approach greatly utilizes the fact that Arabic words are composed of word-parts to improve the efficiency of the recognition framework. The next four sections describe, in more detail, the word-part dictionary, the letter-shape models, the word-part network, and the word recognizer.

5.1. *Word and Word-Part Dictionaries*

To limit the search space, we utilize a dictionary of possible valid words. This ensures better recognition rates compared to systems that can recognize any arbitrary permutation of letters. The Arabic dictionary D is subdivided into a set of sub-dictionaries $D_1, D_2, \cdots, D_n$ based on the number of word-parts in each word. Sub-dictionary D_k includes all words that consist of k word-parts. For example, if a given dictionary D includes the words {انسان, التحدي , ثقافة , جامعة , رواية , فادي , محمد , محمود , معلم , هل , وسام}, D is divided into four sub-dictionaries as shown in Figure 7. We refer to the word-part dictionary $WPD_{k,i}$ as the list of word-parts at the i-th entry (starting from right) of the words in D_k. The word-part dictionaries for D_3 are shown in Figure 8.

5.2. *Letter-Shape Models*

Each Arabic letter has one, two, three, or four shapes that vary according to its position in the word. We have chosen to treat these letter shapes independently

$$
\begin{aligned}
WPD_{3,1} &= \{\text{و , فا , ا}\} \\
WPD_{3,2} &= \{\text{سا, د ,لتحد ,نسا}\} \\
WPD_{3,3} &= \{\text{م, ي ,ن}\}
\end{aligned}
$$

Fig. 8. The word-part dictionaries for D_3 (from Figure 7)

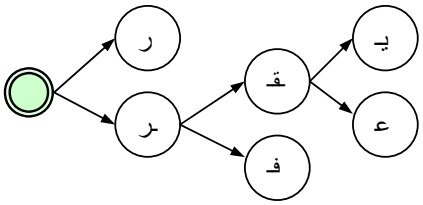

Fig. 9. A word-part dictionary structure, which encodes the word-partsر, فر, عقر, and يقر

(i.e., as unique characters). For example, to each shape of the letter ه(Heh), we associate four letter-shape models ه, هـ, ـهـ, and ـه corresponding to its isolated, initial, medial, and final shape, respectively. The discrete left-to-right HMM without state skipping has been adopted to model each Arabic letter shape. We selected this basic topology because it has been effectively used in handwriting recognition [18]. Additionally, there is no sufficient evidence that more complicated topologies would necessarily achieve better recognition results [18].

5.3. *Word-Part Network*

The letter shapes are embedded in a network that represents the word-part dictionary $WPD_{k,i}$. We optimize this network by grouping all shared suffixes, as shown in Figure 9. Each node in this network represents a letter shape, and each path from the start node (root) to a leaf corresponds to a unique word-part in $WPD_{k,i}$. Each leaf, l, contains the word-part wp_j, which is represented by the path from the start node to the leaf l. We shall refer to this network as a *word-part network* and denote $WPN_{k,i}$ the word-part network that represents the word-part dictionary $WPD_{k,i}$. $WPN^*_{k,i}$ is $WPN_{k,i}$ where each node is replaced with its corresponding letter-shape model. Null transitions are used to connect consecutive letter-shape models in the network as shown in Figure 10.

A word-part network can be constructed by either assigning the first or last letters (of word-parts) to the first level of the tree. Since Arabic word-parts always (except the last word-part in a word) end with one of the six disconnective letters, we assign the last letters to the first level in the word-part network. This fact

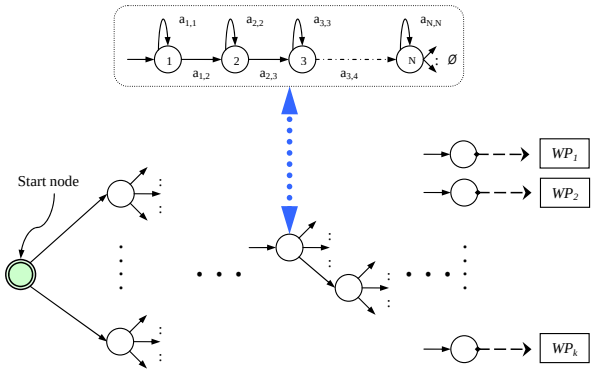

Fig. 10. A word-part network: each path from the start node to a leaf represents a wp_i which is formally defined as $[final + medial^* + initial)|isolated]$.

guarantees that at least one letter is shared in each word-part, which reduces the size of WPN.

5.4. *Arabic Word Recognizer*

In section 4, we discussed the generation of the observation sequences $O_s = [O_1, O_2, \cdots, O_k]$ from a given handwritten Arabic word, where $O_i = [o_{i,1}, o_{i,2}, \cdots, o_{i,T_i}]$ is the observation sequence constructed from the handwritten word-part wp_i. In this section, we introduce our Arabic word recognizer which is based on our word-part network and uses the $Viterbi$ algorithm to determine the recognized word-part.

In a recognition process, we are required to find the word $W = [wp_1, \cdots, wp_k]$, where wp_i is the word-part i, in a given sub-dictionary D_k that maximizes the posterior probability in Equation 1. For simplicity, we assume that all word-parts are statistically independent.

$$P(W|O_s) = \prod^{k} P(wp_i|O_i) \tag{1}$$

where,

$$P(wp_i|O_i) = P(O_i|wp_i)P(wp_i)/P(O_i) \tag{2}$$

We also assume that all word-parts in the sub-dictionary occur with equal probability. As a result, $P(wp)$ is the same for all word-parts and estimating the probability is reduced to maximize $P(O_i|wp_i)$, which can be computed efficiently using the $Viterbi$ algorithm on the given $WPN^*_{k,i}$. The $Viterbi$ algorithm computes $\delta_t(S)$ which refers to the highest likelihood along a single path at time t, that accounts for the first t observation and ends in state S [32]. Specifically, we are only interested

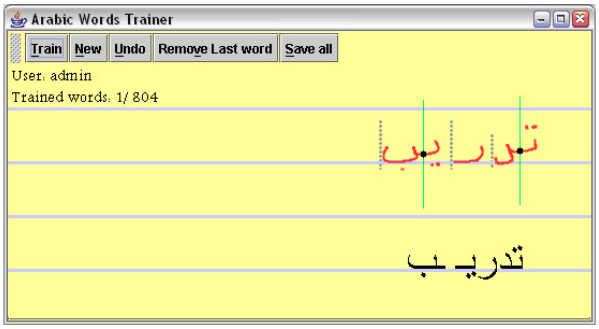

Fig. 11. The trainer is required to manually specify demarcation points that separate letter shapes for letter level training.

in the accumulated maximum likelihood in leaf states at time T_i ($= |O_i|$), given $WPN^*_{k,i}$ and O_i (for $1 < i < k$), which is computed using Equation 3, where q is a leaf state in $WPN^*_{k,i}$, wp is its corresponding word-part, and δ^i_{Ti} is the result of applying the $Viterbi$ algorithm on $WPN^*_{k,i}$. The search for the word W in D_k is performed by applying Equation 4, where W is the recognized word in text format.

$$P(O_i|wp, WPN^*_{k,j}) = \delta^i_{Ti}(q) \tag{3}$$

$$W = \underset{W=[wp_1,wp_2,\ldots,wp_k]\in D_k}{\operatorname{argmax}} \prod_{i=1}^{k} P(O_i|wp_i, WPN^*_{k,i}) \tag{4}$$

6. Model Training

Training data is created by asking Arabic-literate trainers to handwrite (using a digital tablet) a list of predetermined words. The trainers are also asked to manually specify demarcation points, which are points along the word-part curve and a vertical handler, that separate letter shapes such that all delayed strokes of a letter shape are horizontally aligned between the letter shape's demarcation points, as show in Figure 11. As the training process progresses, the system tries to guess the demarcation points based on the accumulated training. Then the trainer can update and correct the demarcation position for incorrect guesses. The details of the specific training data used in our evaluation are discussed in Section 7.

The words in the training data are split into letter-shape samples. In order to avoid improper samples, each letter-shape sample is tested to determine if it satisfies the predetermined letter-shape well-formedness rules, e.g., number and placement of dots/strokes above or below the letter body. The Baum-Welch training algorithm is used to determine the HMM parameters, $\lambda = (A, B, \pi)$, for each letter-shape model. Before the training process, the initial state distribution $\pi = \{\Pi_i\}$ is initialized to: $\pi_1 = 1$ and $\pi_i = 0$ for $1 < i < N$ (where N is the number of states in the model). The transition probability matrix $A = a_{i,j}$ is initialized to $a_{i,i} = 0.5$, and

$a_{i,i+1} = 0.5$ for $i < N$ and $a_{i,j} = 0$ (where, $i \neq j$ and $i \neq j + 1$ for $j < N$, and $a_{N,N} = 1$). The observation matrix B is initialized to reflect a uniform distribution. We have empirically chosen the number of states for each letter-shape model based on the geometric complexity of the letter shape. In our system, the number of states varies from 5 to 11. For example, we assigned 11 states for the isolated letter shape ش(Sheen); and 5 states to the isolated letter shape ا (Alef).

7. Optimization

Time and space complexity play major roles in application efficiency. In interactive applications, such as online script recognition, the system response time is obviously very crucial. Nevertheless, high recognition rates are the most important aspect of these systems.

Segmenting a Latin word into individual letters is an easy task for the non-cursive handwriting and challenging one for cursive writing. In Arabic scripts, such segmentation is often difficult for printed and handwritten words. Several reasons make the segmentation of Arabic words more complicated than cursive Latin words. In Latin cursive writing, the letters are similar to their isolated equivalence and they are usually connected using additional ligatures, which are often not part of the letter body. In addition, Arabic script is not restricted to writing horizontally (along a base line) and allow some letters to appear in vertical orders.

An alternative holistic approach, is the segmentation free scheme. However, these approaches usually require huge databases to store the basic models and high time complexity to search for the right candidates. Each connected component (word-part) is treated as one component to be classified. As a result, a distinct different model is constructed, trained, and saved for each word-part. The number of possible word-parts determines the complexity of the system. Fortunately, in the Arabic language this number is not as huge as one would think. We have processed a collection of Arabic words and found that the number of unique word-parts in Arabic language is around $47,000$ (Table 2) and the majority of these word-parts include four and five letters. These properties are used for further optimizations that include fixing observation-sequence length, and purifying the statistical post-processing phase. Most words in Arabic, more than 90% (as seen in Table 3 and Table 4), have additional strokes and/or loops. Since these features are determined in the feature extraction step, they are utilized to accelerate the classification pro-

cess by reducing the search space. Based on the results in Table 4, an optimization step is performed as a preprocessing step to reduce the number of models to be tested using the number, position, and order of the additional strokes. As shown in Table 4, determining the additional strokes of a written component representing a word-part reduces the size of a class of candidates to less than 500 on average and by that accelerates the system responses, see Table 8.

In our approach, we perform training on the letter level. The models of these letter are combined into a word-part dictionary network, which also represents the models of the word-parts. This network is used to assist and verify word-part recognition and guides combining recognized word-parts into words. Such an approach avoids the training for all valid word-parts in the language, but manage to recognize word-parts and words at high rates even though most of them were not part of the samples training the system.

#Letters	#Word-parts	#Letters	#Word-parts
1	31	6	5532
2	653	7	1253
3	7273	8	198
4	18540	9	19
5	14236	10	3

Table 2. The numbers of valid Arabic word-parts as a function of their size (the number of letters). Note that the one letter length word-parts include few additional letter, such as أ , ة.

DotLoop Property	List of Letters
0 loops	أ , ي ,ن ,ل ,ك ,غ ,ع ,ش ,س ,ز ,ر ,ذ ,د ,خ ,ح ,ج ,ث ,ت ,ب ,ا
1 loop	ص ,ض ,ط ,ظ ,ع ,غ ,ف ,ق ,م ,ه ,و
0 dots	ا ,ح ,د ,ر ,س ,ص ,ط ,ع ,ك ,ل ,م ,ه ,و
1 dot above	ن, ف, غ, ظ, ض,ز, ذ, خ
2 dots above	ة ,ق,ت
3 dots above	ش,ث
1 dots below	ب,ج
2 dots below	ي
optional Loops	ه,ه, ج, ح, خ, ج, ح, خ

Table 3. The number of dots (above/below) and loops define the *DotLoop* property for each Arabic letter. These properties may change as letters change location – beginning, middle, and end – in a word.

Table 4 shows that the word-part dictionary is divided into several disjoint classes based on the number of loops and dots above or below a word-part. Each class has less than 2000 words and the average number is less than 500 words.

To recognize a given word w, we can use an improved algorithm that utilizes the properties of Arabic script.

```
For each w in Text
 For each word-part (wp) in w
   above_dots = CountUpperDots(wp)
   below_dots = CountLowerUpperDots(wp)
   loops = CountLoopsDots(wp)
   ReducedDict = ReduceDictionary(above_dots, below_dots, loops)
   Classify(wp, ReducedDict)
```

To reduce the dictionary search space, we index its words using the number of loops and dots above and below each word. Such indexing reduces the search space to less than 500 word-parts in average, instead of the entire word-part dictionary. In cases were the loop feature is not consistent, the average number would increase to be three times when loops feature is ignored which is still affordable.

Down / Up	0	1	2	3	4	0	1	2	3	4
0	589	746	761	552	242	1154	973	1054	472	195
1	600	623	748	408	192	1520	851	1130	379	192
2	696	597	642	390	151	1693	854	1158	337	191
3	764	546	614	320	128	1563	573	908	217	123
4	499	298	305	99	42	1062	281	585	116	66
	No loops					One loops				
0	813	382	449	119	43	213	53	68	8	6
1	1190	375	673	115	64	361	57	112	9	6
2	1424	425	783	118	84	535	80	204	16	19
3	1078	194	588	59	50	319	34	139	10	12
4	801	117	423	52	37	224	33	87	6	7
	Two loops					Three loops				

Table 4. This table include four matrices that corresponds to 0,1,2, and 3 loops. In each matrix, the cell in column i and row j include the number of word-parts that have i dots above it and j dots below it.

The calculation presented in this paper used around 4 million words from different books and websites. It is obvious that this dataset does not include every word in the Arabic languages. Nevertheless, we believe it is enough to describe the general distribution of Arabic language word-parts.

8. Results and Discussion

Several tests and experiments had been carried out to test and validate the different parts of our system. Four classes of experiments had been performed to measure the effect and validate the following parameters: 1) The datasets for training and evaluation, 2) The proposed feature set, 3) The HMM based classification method, and 4) The optimization process based on additional strokes and loops.

In the contrary to the English case where databases for online handwriting are publicly available for many years, there is no standard reference dataset for training and/or evaluating online handwriting recognition systems for Arabic script. We were not able to access any Arabic handwritten corpus to be used to properly evaluated our approach. Furthermore, it was not possible to attain any online Arabic handwriting recognition system to accurately compare with our results. Therefore, we constructed our own datasets (Manual Database) using four different trainers. Each of the four trainers were guided to write 800 selected words and mark the boundaries of the letter shapes. The words were selected to cover all Arabic letter shapes with almost uniform distribution. In the evaluation stage, ten writers (the four trainers and an additional six new writers) were asked to write 280 words not in the training dataset. The evaluation set included $2,358$ words in total[d]. The overlap of trainers participating in the creation of training and evaluation data is intended to help us evaluate writer-dependence, as well as writer-independence. The trainers and evaluators were asked to write in their own writing style, but respect the rule that a word-part body should be written in a single continuous stroke followed by a number of delayed strokes. We evaluated our system using five different dictionary sizes: $5K$, $10K$, $20K$, $30K$, and $40K$ words selected from the Arabic Treebank [24], twenty random articles from Al-Arabi Magazine, and ten random articles from the website of the news channel Aljazeera. The 280-evaluation words were present in all dictionary sizes. The purpose of the various dictionary sizes is to test our system's performance under different ambiguous conditions.

To validate results using our Manual-Database, We have instructed five students to write many shapes of each letter in all different positions. A novel approach we have developed use these shapes to synthetically generate a compact set of shapes for writing each word-part in the target dictionary. This approach had been used and evaluated in another project and proved to be able to imitate online Arabic handwriting of many different writing styles. Word-part shape generation is performed by concatenating characters in the appropriate position within a word-part for each word. Since the number of different shapes for each word is very large, this number is reduced using dimensionality reduction and clustering techniques. We used this system to synthetically generate all the manually generated words, which was used to training and test the system. Table 5 shows the recognition rates

[d]Not all volunteers finished the testing task, and some word samples were omitted due to being incomplete.

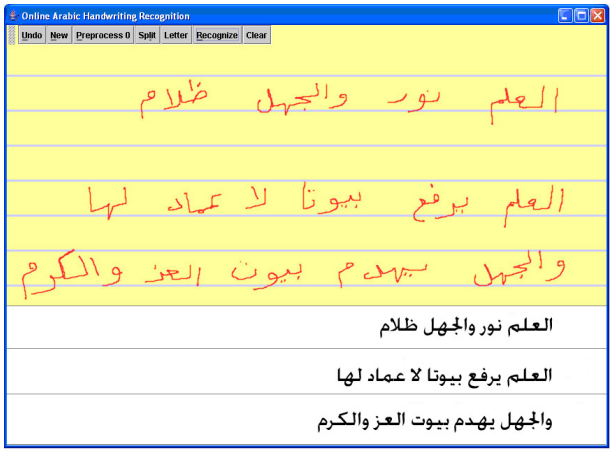

Fig. 12. Several handwritten sentences (top) and their correct recognition (bottom).

of our system using our Manual-Database and the Synthetic-Database for training and testing.

Database Size		5K	10K	20K	30K	40K
Manual Database	WD	98.44%	97.94%	96.86%	95.90%	95.44%
	WI	98.49%	97.78%	96.54%	95.12%	94.44%
Synthetic Database	WD	92.14%	91.15%	89.61%	88.33%	85.17%
	WI	91.44%	91.01%	89.64%	88.22%	88.11%

Table 5. Writer-dependent (WD) and writer-independent (WI) average word recognition rates for two tests including 2,358 and 6220 words written by ten writers. Results using the synthetic database to extract the same words as in the manual database are in the lines three and four.

To evaluate the effectiveness of the presented set of features we carried out several experiments, while keeping the same system and training data sets but replacing the geometric features set by a set of sliding-window based features. Pechwitz and Maergner [31] used the sliding-window features to recognize written Arabic words using the IFN/ENIT benchmark database. To adjust the online strokes to the offline case we thicken the one pixel width stroke to three pixels width before applying the sliding window technique. We extract the features from the binary image instead of the gray one. The feature vector is the concatenation of the feature vectors extracted form the three sliding windows. Table 6 shows that recognition rates in both cases are similar.

We designed a system based on Dynamic Time Warping (DTW) technique for the matching and classification process. This system uses the same feature set

Database Size		5K	10K	20K	30K	40K
Geometric Features	WD	98.44%	97.94%	96.86%	95.90%	95.44%
	WI	98.49%	97.78%	96.54%	95.12%	94.44%
Sliding Win Features	WD	91.21%	91.15%	90.61%	88.63%	88.21%
	WI	92.11%	91.31%	90.22%	90.11%	86.78%

Table 6. Writer-dependent (WD) and writer-independent (WI) average word recognition rates using our geometric features and pixel based features from the sliding window technique.

extracted from the word-parts to build the collection of prototypes for matching. The words from the training sets were used to build the sets of prototypes and datasets of feature vectors were saved for each word instead of the HMM model. Table 7 shows the results of the two classifiers.

Database Size		5K	10K	20K	30K	40K
HMM classifier	WD	98.44%	97.94%	96.86%	95.90%	95.44%
	WI	98.49%	97.78%	96.54%	95.12%	94.44%
DTW classifier	WD	91.24%	90.21%	90.12%	89.23%	88.27%
	WI	96.18%	96.11%	93.32%	90.18%	87.22%

Table 7. Results of our system compared a system with the same database and feature sets but using a different classifier based on DTW matching technique.

High recognition rates are the most important aspect in our work, even though the system response time is obviously very important. In these experiments we intend to show the applicability of the suggested optimization to improve recognition rates and reduce the time response.

Dictionary Size		5K	20K	40K
Writer Independent	Time response Reduction	-62.11%	-75.31%	-78.45%
	Improvement Recognition	+0.91%	+1.63%	2.87%
Writer Dependent	Time response Reduction	-64.33%	-76.18%	-78.98%
	Improvement Recognition	+1.12%	+2.32%	3.14%

Table 8. The improvement in response time and recognition rates that results from using the dots and loops for optimization.

Reducing time response was not the main task in this experimental test. Therefore, results are presented as improvement percentage to validate the effectiveness of the optimization independent from the time efficiency of the system. Table 8 shows that using additional strokes combined with loop counter in a preprocessing

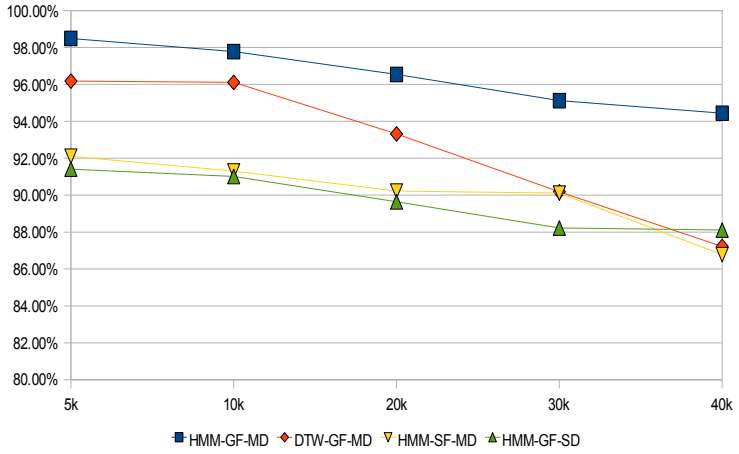

Fig. 13. This graph shows the results of recognition rates comparing the different systems. The compared systems are our proposed system and other three systems changing one factor each time. The factors we changed are: SD = Synthetic Database, SF = Sliding Window features, and DTW = Dynamic Time Warping classifier.

can reduce the search space and improve response time. It is obvious that this factor becomes more important when the target database is large.

Overall, we achieved good results given that we used a relatively small training set. The differences between the writer-independent and writer-dependent recognition rates are less than 2%, with all tested dictionary sizes. This implies that the features, model, and delayed-stroke algorithm, we introduced, are adequate for writer-independent handwriting recognition. The performance degrades as the dictionary size increases. The degradation in word-part recognition is at a lower rate than word recognition, suggesting that the recognition failure is tied to specific word-parts. Most of the recognition errors arose in word-parts that have similar shapes, such as ب/بـ and د/ر. Therefore, the current features are not sufficient for adequately distinguishing between such word-parts.

9. Conclusion and Future Work

This paper introduced an HMM-based system with novel components to provide solutions for most of the inherent difficulties in recognizing Arabic script – letter connectivity, position-dependent letter shaping, and delayed strokes. An evaluation of the system shows that the used features and letter models are adequate for

writer-independent handwriting recognition at high rates. Our solution for delayed strokes can also be utilized to recognize scripts that include diacritical marks (e.g., French, German, Spanish, etc.).

In the future, we plan to increase the system's robustness to handle cases where delayed strokes are written before the completion of a word-part. We also plan to reduce the number of errors described in Section 8, using geometric-computation techniques and a more sophisticated post-processing phase. Moreover, we plan on exploring sentence-level language modeling to improve word recognition [26].

References

1. S. Al-Emami and M. Usher. On-line recognition of handwritten arabic characters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):704–710, 1990.
2. R. Al-Hajj, C. Mokbel, and L. Likforman-Sulem. Combination of hmm-based classifiers for the recognition of arabic handwritten words. In *ICDAR 2007. Ninth International Conference on Document Analysis and Recognition*, volume 2, pages 959 – 963, Sept 2007.
3. H. A. Al-Muhtaseb, S. A. Mahmoud, and R. S. Qahwaji. Recognition of off-line printed arabic text using hidden markov models. *Signal Process.*, 88(12):2902–2912, 2008.
4. A. T. AL-Taani. An efficient feature extraction algorithm for the recognition of handwritten arabic digits. *International journal of computational intelligence*, 2(2), 2005.
5. A. M. Alimi and O. A. Ghorbel. The analysis of error in an on-line recognition system of arabic handwritten characters. In *ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 2)*, page 890, Washington, DC, USA, 1995. IEEE Computer Society.
6. A.l M. Alimi. An evolutionary neuro-fuzzy approach to recognize on-line arabic handwriting. *icdar*, 00:382, 1997.
7. B. Alsallakh and H. Safadi. Arapen: An arabic online handwriting recognition system. In *Information and Communication Technologies, 2006. ICTTA '06. 2nd*, volume 1, pages 1844– 1849, April 2006.
8. A. Amin. Machine recognition of handwritten arabic word by the irac ii system. In *In Proceedings of the 7th Joint on Pattern Recognition*, pages 35–37, October 1982.
9. M. S. Baghshah, S. B. Shouraki, and S. Kasaei. A novel fuzzy approach to recognition of online persian handwriting. In *ISDA '05: Proceedings of the 5th International Conference on Intelligent Systems Design and Applications*, pages 268–273, Washington, DC, USA, 2005. IEEE Computer Society.
10. H. Beigi, K. Nathan, G. Clary, and J. Subrahmonia. Size normalization in online unconstrained handwriting recognition. In *The IEEE International Conference on Image Processing*, volume I, pages 169–172, November 13-16 1994.
11. A. Benouareth, A. Ennaji, and M. Sellami. Arabic handwritten word recognition using hmms with explicit state duration. *EURASIP Journal on Advances in Signal Processing*, 2008:13, 2008.
12. Fadi Biadsy. Online arabic handwriting recognition. M.Sc Thesis, Ben Gurion University of the Negev, 2005.
13. Fadi Biadsy, Jihad El-Sana, and Nizar Habash. Online arabic handwriting recognition using hidden markov models. In *IWFHR '10 2006, France*, 2006.
14. M. Dehghana, K. Faeza, M. Ahmadi, and M. Shridhar. Handwritten farsi (arabic) word recognition: a holistic approach using discrete hmm. *Pattern Recognition*,

34(5):1057–1065, May 2001.
15. D. Douglas and T. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10(2):112–122, 1973.
16. T. S. El-Sheikh and S. G. El-Taweel. Real-time arabic handwritten character recognition. *Pattern Recogn.*, 23(12):1323–1332, 1990.
17. R. Halavati, M. Jamzad, and M. Soleymani. A novel approach to persian online hand writing recognition. *Transactions ON Engineering and Technology*, 6(1305-5313.), June 2005.
18. J. Hu, S. G. Lim, and M. K. Brown. Writer independent on-line handwriting recognition using an hmm approach. *Pattern Recognition*, 33(1):133–147, 2000.
19. J. Hu, S. C. Oh, J. H. Kim, and Y.B. Kwon. Unconstrained handwritten word recognition with interconnected hidden markov models. In *In proceedings Third Int. Workshop on Frontiers in Handwriting Recognition*, pages 455–560, 1993.
20. M. S. Khorsheed. *Automatic recognition of words in arabic manuscripts*. PhD thesis, University of Cambridge, 2000.
21. M. S. Khorsheed. Recognising handwritten arabic manuscripts using a single hidden markov model. *Pattern Recogn. Lett.*, 24(14):2235–2242, 2003.
22. M. S. Khorsheed. Offline recognition of omnifont arabic text using the hmm toolkit (htk). *Pattern Recogn. Lett.*, 28(12):1563–1571, 2007.
23. J. J. Lee, J. Kim, Jin, and H. Kim. Data driven design of hmm topology for online handwriting recognition. In *in The 7th International Workshop on Frontiers in Handwriting Recognition*, pages 107–121. World Scientific Publishing Company, 2001.
24. M. Maamouri. Developing an arabic treebank: Methods, guidelines, procedures, and tools. In *In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (COLING*, 2004.
25. S.i Mahmoud. Recognition of writer-independent off-line handwritten arabic (indian) numerals using hidden markov models. *Signal Process.*, 88(4):844–857, 2008.
26. J. Makhoul, T. Starner, R. Schwartz, and G. Chou. On-line cursive handwriting recognition using speech recognition methods. In *Proceeding of IEEE ICASSP'94*, pages V125–V128. IEEE, April 1994.
27. F. Menasri, N. Vincent, M. Cheriet, and E. Augustin. Shape-based alphabet for off-line arabic handwriting recognition. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 969–973, Washington, DC, USA, 2007. IEEE Computer Society.
28. N. Mezghani, M. Cheriet, and A. Mitiche. Combination of pruned kohonen maps for on-line arabic characters recognition. In *ICDAR '03: Proceedings of the Seventh International Conference on Document Analysis and Recognition*, page 900, Washington, DC, USA, 2003. IEEE Computer Society.
29. N. Mezghani, A. Mitiche, and M. Cheriet. On-line recognition of handwritten arabic characters using a kohonen neural network. In *IWFHR '02: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02)*, page 490, Washington, DC, USA, 2002. IEEE Computer Society.
30. S.C. Oh, J.Y. Ha, and J.H. Kim. Context-dependent search in interconnected hidden markov model for unconstrained handwriting recognition. *Pattern Recogn.*, 28(11):1693–1704, November 1995.
31. Mario Pechwitz and Volker Maergner. Hmm based approach for handwritten arabic word recognition using the ifn/enit- database. In *ICDAR '03: Proceedings of the Seventh International Conference on Document Analysis and Recognition*, page 890, Washington, DC, USA, 2003. IEEE Computer Society.

32. L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Readings in speech recognition*, pages 267–296, 1990.
33. S. M. Razavi and E. Kabir. A data base for online persian handwritten recognition. In *6th Conference on Intelligent Systems, In Farsi*, 2004.
34. L. Schomaker. Using stroke or character-based self-organizing maps in the recognition of on-line, connected cursive script. *Pattern Recognition*, 26(3):443–450, March 1993.